"*Covid vaccine* is against Covid but *Oxford vaccine* is made at Oxford!" Semantic Interpretation of Proper Noun Compounds

Keshav Kolluru[†] **Gabriel Stanovsky**[‡] **Mausam**[†]
[†] Indian Institute of Technology Delhi
keshav.kolluru@gmail.com, mausam@cse.iitd.ac.in
[‡] The Hebrew University of Jerusalem
gabriel.stanovsky@mail.huji.ac.il

Abstract

Proper noun compounds, e.g., "*Covid vaccine*", convey information in a succinct manner (a "Covid vaccine" is a "vaccine that *immunizes against* the Covid *disease*"). These are commonly used in short-form domains, such as news headlines, but are largely ignored in information-seeking applications. To address this limitation, we release a new manually annotated dataset, PROPNCI, consisting of 22.5K proper noun compounds along with their free-form semantic interpretations. PROPNCI is 60 times larger than prior noun compound datasets and also includes non-compositional examples, which have not been previously explored. We experiment with various neural models for automatically generating the semantic interpretations from proper noun compounds, ranging from few-shot prompting to supervised learning, with varying degrees of knowledge about the constituent nouns. We find that adding targeted knowledge, particularly about the common noun, results in performance gains of upto 2.8%. Finally, we integrate our model generated interpretations with an existing Open IE system and observe an 7.5% increase in yield at a precision of 85%. The dataset and code are available at https://github.com/dair-iitd/pronci.

1 Introduction

Proper noun compounds (PNCs) (Breban et al., 2019)[1] are grammatical constructions where a proper noun is followed by a common noun, for example: *Covid vaccines* or *Buddhist monks*. These often serve as a compact way to convey information about an already known entity, omitting predicates that are interpreted by the reader using surrounding context, common sense, and world knowledge. For example, a reader is likely to interpret that "Buddhist monks" are "*religious people who are* buddhists". In other cases, PNCs are used to identify specific entities, and do not provide additional information. For example, *Watergate scandal* and *Kawasaki disease* do not have any implicit relation between the proper and common noun as they refer to a specific instance of a scandal and a disease. Table 1 provides additional examples.

Thanks to their brevity, PNCs are commonly used to shorten descriptions in space-constrained domains, such as news articles headlines (Breban et al., 2019). However, we find that prior work on compound noun interpretations only considered cases where the constituents are common nouns (e.g. *baby oil*), thus missing all of the information conveyed in proper noun compounds (Shwartz and Waterson, 2018; Hendrickx et al., 2013).

To address this limitation in current systems, we begin by defining the task of PNC interpretation as two subsequent stages (Section 3). The first stage requires identifying whether a given PNC is compositional or not, while the second stage is the generation of an interpretation, where applicable.

In Section 4, we present PROPNCI, a crowdsourced dataset over Wikipedia containing 22.5K proper noun compounds and their annotated semantic interpretations. Candidates PNCs are found using syntactic parsing, and are then presented to crowdworkers who are asked to interpret them. Our annotation interface marks whether workers needed to read the full sentence, thus identifying PNCs whose interpretation relies on context. We will make the PROPNCI dataset publicly available to spur future research into PNCs.

In Section 5, we develop two approaches for PNC interpretation: (1) a multi-task neural model that performs classification and sequence generation in two distinct stages and (2) a text-to-text approach, using a sequence-to-sequence model for both classification and generation. In addition, we experiment with different methods for injecting various sources of world knowledge, which seems crucial for the task, using external resources like Wikipedia and WordNet (Fellbaum, 2010), that

[1] also referred to as proper noun modified compounds.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10407–10420
December 7-11, 2022

Type	Example	Semantic Interpretations
Proper NC (*Proper-Common*)	*Shakespeare biography*	is a biography about Shakespeare
	London theatre	is a theatre in London *;* is a theatre located in London
	Concorde airplane	[NON-CMP] (Non-Compositional)
	Notre-Dame cathedral	[NON-CMP] (Non-Compositional)
Common NC (*Common-Common*)	*nursing job*	is a job in nursing field *;* is a job involving nursing
	oil price	is price paid for the oil

Table 1: Examples of common and proper noun compounds along with their semantic interpretations ("*;*" separates multiple interpretations). [NON-CMP] indicates the absence of implicit relation between the constituent nouns.

give relevant information or definitions about the PNCs, that help in improving performance.

For evaluating the generated interpretations, we propose a combination of classification-based metric and generation metrics to properly handle both the interpretable and non-interpretable cases, respectively (Section 6). Since multiple correct interpretations are possible for a PNC, we use learned metrics such as BLEURT (Sellam et al., 2020), that is finetuned on human-annotated preferences.

Finally, we show that training on PRONCI yields models that can readily benefit extrinsic downstream application in the task of Open Information Extraction (Banko et al., 2007), thus widely extending their coverage (Section 8). Our approach first automatically extracts PNC interpretations using our models, then introduces it explicitly back into an Open IE extraction using a sequence to sequence model, thus giving an interpretation-integrated extraction. We then apply a high precision rule to generate new relations which leads to a 7.5% increase in yield at an estimated precision of 85% on the added extractions, when compared to extractions generated from the original sentences themselves. A major advantage of this approach is that it is agnostic to the Open IE system being used. To conclude, our main contributions are:

1. We introduce the PRONCI dataset, containing interpretation for 22.5K proper noun compounds and their semantic interpretations.

2. We develop multi-task and generation based neural baselines that can leverage external knowledge for achieving higher performance.

3. We design metrics for evaluating the quality of generated semantic interpretations.

4. We demonstrate the usefulness of the generated interpretations in a downstream application by using them to augment the expressivity of Open IE systems.

2 Related Work

Noun compounds are commonly used in English language, constituting 3.9% of the tokens in the Reuters corpus (Baldwin and Tanaka, 2004). They can be arbitrary length phrases, such as *split air conditioner*, but most prior work on interpreting noun compounds has primarily looked at two word noun compounds of the type *noun-noun*, where both are common nouns. To the best of our knowledge challenges in interpretation where the first word is a proper noun (i.e., *proper noun compounds*) have not been addressed, although their functional analysis and prevalence in certain domains have been studied in linguistics (Rosenbach, 2007; Alexiadou, 2019; Breban et al., 2019). We briefly summarise the various types of noun-compound interpretations in literature and discuss their uses in applications.

Types of interpretation: Various types of interpretations for noun compounds have been explored, covering classification, ranking and generation. Prior literature has frequently posed the interpretation as a **classification** task, where the classes can belong to abstract labels (Fares, 2016), semantic frame elements (Ponkiya et al., 2018) or prepositions (Lauer, 1995) However, none of these schemes can cover all range of possible noun compounds, thus limiting their expressivity and coverage. SemEval 2010 Task 9 (Butnariu et al., 2009) annotates human preferences for a set of 25-30 templatized paraphrases for each of the 250 training and 300 testing noun compounds. The task is framed as producing an accurate score for each paraphrase that **ranks** them in the correct order. SemEval 2013 Task 4 (Hendrickx et al., 2013) released a dataset of noun compounds and annotated free paraphrases for each compound. Participating models were evaluated by matching and scoring the **generated** predictions with the gold set.

Ponkiya et al. (2020) is the current state of art which poses the problem as generation of masked tokens using a pretrained T5 model (Raffel et al., 2020) to get free paraphrase interpretations in a completely unsupervised manner. This leads to better performance than techniques that use the available training data. However, with the PRONCI dataset, we do find that supervised models do outperform zero-shot models due to the scale.

Applications: Noun compound interpretations have been helpful in translation of noun compounds by either using a one-to-one mapping of interpreted prepositions (Paul et al., 2010) or using recursive translation patterns (Balyan and Chatterjee, 2015). In Question Answering systems, they have been used for disambiguating different types of noun-noun compounds in passage analysis (Ahn et al., 2005). They have also been useful for normalizing text that can help textual entailment (Nakov, 2013) and as auxiliary semantic annotation modules to improve parsing (Tratz, 2011). In this work, we show their use in the task of Open IE.

Open Information Extraction (Open IE) (Banko et al., 2007; Mausam, 2016; Kolluru et al., 2020b) involves extracting a set of tuples from the sentence where each field of the tuple contains phrases from the sentence itself. This makes it ontology-agnostic and allows it to be used for creation of domain agnostic Open Knowledge Bases (Broscheit et al., 2020; Vashishth et al., 2018; Gupta et al., 2019). The relations are often verb-based (Fader et al., 2011) but can also be noun-mediated (Pal and Mausam, 2016) or involve implicit information (Soderland et al., 2015).

Fader et al. (2011) relied on high precision rules to extract a wide variety of verb-mediated relations. Soderland et al. (2015) uses dependency paths for generating high precision extractions based on three implicit relations, *has job title*, *has city* and *has nationality*. Pal and Mausam (2016) considers noun mediated relations that can be extracted from compound noun phrases while dealing with challenges involved with denonyms and compound relational nouns. However, none of them consider implicit relations present in noun compounds.

Moreover, recent state of art Open IE systems like OpenIE6 (Kolluru et al., 2020a) and Gen2OIE (Kolluru et al., 2022) rely on bootstrapped examples (generated using OpenIE4 (Pal and Mausam, 2016; Christensen et al., 2011)) for training. Therefore they only generate extractions that contain phrases from the text and miss the cases where the content words are implicit. OpenIE6 (Kolluru et al., 2020a) adopts a pipeline approach to integrate conjunction splitting into Open IE outputs, where coordination analysis and sentence splitting is performed as a preprocessing step, and the Open IE extractions are generated from the split sentences which are then merged.

Task Instructions
1. Your goal is to describe the relation between the two words by filling in the blanks. 2. You can write up to five words (or less!) 3. The resulting relation should form a valid English sentence (see below for an example). 4. You can consult an example sentence as additional context, but the relation you write should be inferred only from the two words, and not by additional information. 5. If it is a name, entity, location or if you can't describe the relation between the words, please leave the relation blank.
Examples
1. Coke Spokesman *is a worker of* Coke. 2. Leake government *is located in* Leake. 3. Capitol Hill
Pitfalls
1. Coke Spokesman *employment* Coke. The relation should form a valid sentence. 2. Leake government *has a failed* goverment. The relation should be inferred by the words themselves and not by additional context.

Table 2: Instructions for the task along with examples and common pitfalls that are provided to the human workers from AMT for constructing PRONCI dataset.

3 Problem Definition

Interpretations of noun compounds are meant to expose the expressed implicit relation. Free-form paraphrases as interpretations provide flexibility for expressing relations implied in noun compounds, overcoming the limitations associated with choosing from a fixed set of classes or templates at the cost of a possibly non-consolidated representation, i.e., where similar-meaning noun compounds are represented differently. Hence, we define semantic interpretation of a PNC as a free-form paraphrase that exposes the implicit relation between the constituent nouns, if any relation exists, else identify it as non-compositional ([NON-CMP]).

$$\text{SemInt}(pnc) = \begin{cases} \text{Paraphrase}, & \text{if } \textit{reln.} \text{ exists} \\ [\text{NON-CMP}], & \text{if } \textit{reln.} \text{ absent} \end{cases}$$

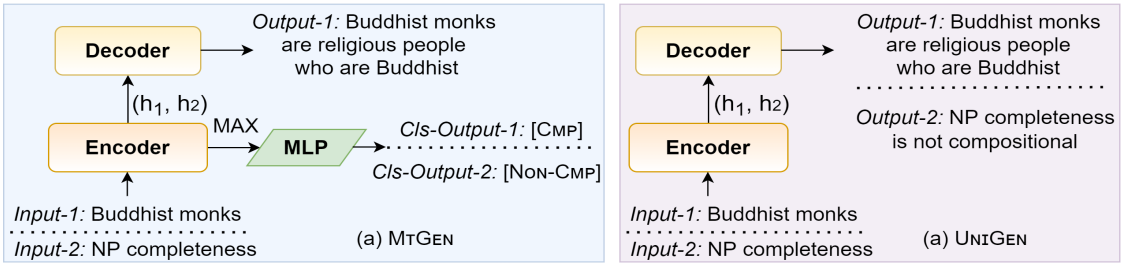

Figure 1: MTGEN, a multi-task Seq2Seq model classifies the example into (non) compositional classes and generates the interpretation where valid, while the UNIGEN, an unified generation model, uses a Seq2Seq model to generate interpretations or identify non-compositional examples using a specific string "is not compositional".

4 PRONCI Dataset

To facilitate research in semantic understanding of proper noun compounds, we collect and release a supervised dataset as part of this work, which we call the PRONCI dataset. It contains 22,500 PNCs and their semantic interpretations which are annotated by human workers hired from Amazon Mechanical Turk (AMT).

The scale of the dataset is orders of magnitude greater than previously published free-paraphrase (common) noun compound datasets like SemEval 2013 Task 4 (Hendrickx et al., 2013) that have only considered 355 noun compounds. For handling the evaluation of generated interpretations where multiple correct answers are possible, prior datasets choose to annotate multiple interpretations for each noun compound (varying from 30-50). On the other hand, PRONCI dataset only contains one interpretation per noun compound as we choose to invest our annotation budget in breadth rather than depth, relying on recent advances in semantic text similarity (e.g. BLEURT (Sellam et al., 2020)) to help evaluate the generated interpretations.

Moreover, prior datasets consider noun compounds out of context, while PRONCI also contains the sentence in which the proper noun compound is used. Providing this additional context helps to limit the ambiguity associated with multiple possible interpretations of the noun compound. For example, *U.S. sanctions* can mean either sanctions imposed by U.S. or sanctions imposed on U.S. The exact case can be determined based on the context in which it is used. "*U.S. sanctions* on Iran have crippled the country", implies the former and "*U.S. sanctions* by Iran..." implies the latter.

To prepare the PRONCI dataset, we randomly sample sentences from Wikipedia, and retain sentences which contain two-word proper noun compounds as identified by the SpaCy dependency parser (Honnibal et al., 2019). For every word, SpaCy identifies the root word along with the dependency tag. The "compound" dependency tag is used if the word and its root are part of a compound word. Then the parts of speech of the first and second word of the compound are checked. If they are proper noun ("PROPN") and common noun ("NOUN") respectively, we identify it as a proper noun compound and include it. If any word pairs have been identified incorrectly as proper noun compounds, they are marked by annotators to indicate the absence of any relation.

After the collection of proper noun compounds and corresponding sentences in which they appear, we posted HITs on the AMT platform for identification of relation between two words. The HITs were accompanied by task instructions, summarized in Table 2. The workers were paid 9 USD per hour on average, based on initial annotation experiments which indicated an average annotation time of 20 seconds on each compound.

To check the quality of annotation, we randomly sample 100 examples and find them to be correct 93% of the time. This represents an acceptable level, considering the difficulty of understanding certain compounds that need technical knowledge (*AES key*) or cultural background (*Abner characters*), as well as the subjectiveness in determining non-compositionality.

5 Models

The task of semantic interpretation of proper noun compounds involves generating valid paraphrases that explicate the relation in cases which are compositional. So a model designed for this task needs to first identify if the given noun compound is compositional ([CMP]) or not ([NON-CMP]), and generate a paraphrase accordingly. We experiment with (1) supervised neural models, (2) adding external infor-

Knowledge	Example
None	Buddhist monks
Sentence	Recent visitors to the campus include Buddhist monks who installed an environmental artwork at Lower Pond. [*SEP*] Buddhist monks
WordNet-NN	Buddhist meaning: Buddhism is a widespread Asian religion based on a series of original teachings attributed to Gautama Buddha. [*SEP*] Buddhist monks
Wiki-NNP	monks meaning: a male religious living in a cloister and devoting himself to contemplation and prayer and work [*SEP*] Buddhist monks
NER-NNP	Buddhist belongs to nationalities or religious groups [*SEP*] Buddhist monks

Table 3: Examples demonstrating the addition of different sources of knowledge for the compound, "Buddhist monks", in form of prompts that are concatenated with [*SEP*] token. NNP and NN correspond for information about proper and common noun respectively, which can be from WordNet, Named Entity tags or Wikipedia.

mation and (3) zero/few-shot prompting models.

Supervised neural models: We use two types of supervised neural models: (1) a multi-task and (2) a unified generative model. Both models are depicted in Figure 1. The multi-task neural model uses a single model to perform both the tasks of classification as well as generation. For classification, the model uses the max-pooled representations of encoder hidden states that is passed to an MLP (Maini et al., 2020) to get the corresponding class probabilities of [CMP] and [NON-CMP]. In case the example is classified as compositional, a decoder is used for generating the paraphrase. We refer to this model as MTGEN.

In the unified generation model, we follow the recent advances in NLP where multiple tasks are posed in a common text-to-text format and are handled by a single Seq2Seq model like T5 (Raffel et al., 2020). For this purpose, we pose the task as a simple string generation problem that outputs either the paraphrase itself in cases where it is interpretable or generates the string "*proper noun compound* is non-compositional" in the remaining cases. We refer to this model as UNIGEN.

External information: Since the task of interpretation requires knowledge of the noun compound, we also experiment with adding different types of knowledge to the model that help it in generating accurate interpretations. Various methods have been proposed to incorporate external knowledge into pre-trained language models (Wang et al., 2021; Liu et al., 2022b; Verga et al., 2021). We use a simple strategy of concatenating the knowledge along with the proper noun compound before passing it to the model. A [*SEP*] token is added as a demarcator to differentiate the added knowledge.

We use four sources of knowledge that provide further information about the noun compound. They include information of the proper noun, from (1) the first paragraph of Wikipedia that an entity linking system links it to (Wiki-NNP), (2) tags assigned to it by the Named Entity Recognition system (NER-NNP), or include information about the common noun using (3) the corresponding synset definitions provided in Wordnet (WordNet-NN), or information about the entire compound based on the (4) sentence in which it is used. An example of each type of knowledge is shown in Table 3.

Zero/Few-shot prompting: Prior techniques for noun compound interpretation such as (Ponkiya et al., 2020) have proposed zero-shot generation using pre-trained language models to achieve state-of-art performance on SemEval 2013 Task 4 (Hendrickx et al., 2013) and SemEval 2010 Task 9 (Butnariu et al., 2009). We therefore evaluate the performance of such techniques along with some extensions using few-shot learning on the PRONCI dataset. We find that there exists a significant gap compared to finetuning on the supervised dataset, demonstrating the importance of having a large scale dataset for the task of PNC interpretation.

6 Experimental Setup

The 22,500 examples of PRONCI are split into train, validation and test such that all compounds with the same common noun occur exclusively in a single set. Such splitting ensures that there is no intersecting common noun in either the train or evaluation splits. This results in a more challenging setting than splitting the examples randomly, whose results are shown in Appendix B. Further, we also consider subsets that contain only compositional examples (CMP) or only non-compositional examples (Non-CMP). The number of examples in each case are shown in Table 4.

The dataset has 7,383 unique relations, with ev-

Type	#Train	#Validation	#Test	#Total
CMP	9,722	1,416	2,497	14,389
Non-CMP	5,568	934	1,609	8,111
All	15,290	2,350	4,106	22,500

Table 4: Number of training, validation and testing examples in the PRONCI dataset. CMP indicates the subset that contains only compositional examples and constitute 63.9% of the examples. Non-CMP indicates the complementary subset that contains only non-compositional examples and constitute the remaining 36.1% of the examples.

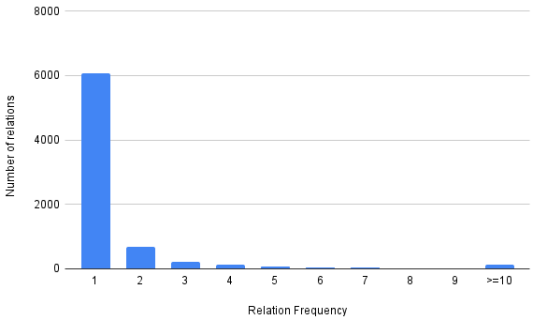

Figure 2: Plot of relation distribution in the PRONCI dataset. It shows the number of relations that have a frequency of 1 to 9 and >=10.

ery relation occurring in an average of 1.84 examples. It contains 6,061 relations that occur only once in the dataset, as shown in Figure 2. The top 5 most commonly occurring relations along with their frequency (indicated in brackets) are *is located in* (560), *is based in* (389), *are relatives of* (245), *is an area of* (215) and *are located in* (125).

Evaluation metrics: Since the task involves a combination of classification and generation, the evaluation metric uses either exact match or semantic similarity depending on the type of example. If an example has either the model prediction (**p**) or the gold annotation (**g**) as non-compositional, then an exact match (EX-MATCH) between the prediction and gold gives a binary score of 0 or 1. In examples where both the gold annotation and model prediction are compositional, a semantic matching algorithm (SEM-MATCH) is used to give a score between 0 and 1 which indicates the extent of their similarity.

$$\text{Score}(\mathbf{g}, \mathbf{p}) = \begin{cases} \text{SEM-MATCH}(\mathbf{g}, \mathbf{p}), \text{if CMP} \\ \text{EX-MATCH}(\mathbf{g}, \mathbf{p}), \text{if Non-CMP.} \end{cases}$$

In particular, we compare two alternatives for SEM-MATCH: (1) the rule-based popular BLEU score, relies on *n-gram* overlap, and often used in machine translations; and (2) BLEURT, which is finetuned over pretrained language model and represents a recent trend in trained evaluation metrics for text generation tasks.

In both alternatives, we use the entire paraphrase to evaluate the semantic score as evaluating only the relations does not suit metrics such as BLEURT, which expects a full-formed sentence to infer the semantic meaning. Evaluation of the quality of BLUERT for similarity between predicted and gold paraphrases using 1K human annotated judgements indicates a 0.57 Pearson and 0.56 Kendall correlation. We follow standard protocols in evaluating metric quality, as used in WMT Metrics shared tasks, and ask human annotators to rate the compositional model predictions as good, average and bad and see how these judgement scores correlate with the BLEU and BLEURT scores. Further details are provided in Appendix A.

We denote the final evaluation metric as SEM/EX-MATCH. When using BLEU or BLEURT as the semantic matcher, the metric is also referred to as BLEU/EX or BLEURT/EX, respectively. To understand the effect of each type of match, we also report the EX-MATCH classification accuracy over all the examples, where the compositional type is assigned the positive class, and the non-compositional type is assigned as the negative class. Along with binary accuracy, we compute the precision and recall as well. Since the SEM-MATCH cannot be computed over all examples, we report the scores averaged over only the cases where both gold and prediction are compositional.

Pre-trained models: For all our experiments, we use the T5-base (Raffel et al., 2020) as the default initialization, unless explicitly mentioned otherwise. It contains 220M parameters. For checking the statistical consistency, every model is trained 5 times with different seeds and their mean and standard deviation are reported.

Hyper-parameters and computational resources: We run all our experiments using a V100 GPU. We use the standard hyper-parameters recommended in T5 for all the experiments, using a batch size of 16, initial learning rate of 2e-5. The final model is chosen using early stopping on the validation set after training for 10 epochs. Each round of training and evaluation takes around 1 hr.

Model	Knowledge	EX-MATCH			SEM-MATCH		SEM/EX-MATCH	
		Precision	Recall	Accuracy	BLEU	BLEURT	BLEU	BLEURT
MTGEN	None	**79.1** ± 1.37	67.1 ± 1.84	**79.5** ± 0.58	32.7 ± 1.61	57.9 ± 0.42	44.3 ± 1.05	57.5 ± 0.66
	Sentence	78.1 ± 1.51	68.4 ± 2.50	79.4 ± 0.25	34.7 ± 0.36	58.3 ± 0.76	45.7 ± 0.60	57.8 ± 0.75
	WordNet-NN	74.2 ± 3.71	76.4 ± 5.68	79.4 ± 1.08	33.2 ± 1.08	57.6 ± 0.51	47.1 ± 0.92	58.9 ± 0.76
	Wiki-NNP	52.8 ± 2.43	**90.6** ± 3.02	63.2 ± 2.96	24.0 ± 0.36	32.9 ± 2.38	43.0 ± 0.50	45.4 ± 0.98
	NER-NNP	**79.1** ± 0.63	67.7 ± 1.63	79.7 ± 0.55	34.5 ± 0.23	59.2 ± 0.37	45.4 ± 0.51	58.3 ± 0.68
UNIGEN	None	73.5 ± 2.99	74.4 ± 2.26	78.7 ± 1.40	34.1 ± 1.99	58.6 ± 0.78	46.7 ± 1.12	58.6 ± 0.94
	Sentence	73.0 ± 1.57	77.6 ± 1.83	79.3 ± 0.55	34.4 ± 0.81	58.8 ± 0.68	**47.9** ± 0.41	**59.5** ± 0.57
	WordNet-NN	65.3 ± 5.76	82.9 ± 5.05	74.5 ± 3.74	33.7 ± 0.88	56.5 ± 0.65	47.4 ± 0.45	56.7 ± 1.52
	Wiki-NNP	65.3 ± 3.05	66.3 ± 5.50	71.8 ± 1.32	25.7 ± 0.59	37.8 ± 2.13	38.4 ± 1.55	43.9 ± 1.09
	NER-NNP	75.7 ± 0.95	72.3 ± 1.52	79.4 ± 0.21	**35.2** ± 0.23	**59.4** ± 0.40	46.9 ± 0.45	59.0 ± 0.42

Table 5: Performance of MTGEN and UNIGEN on the PRONCI dataset trained under five different knowledge settings. All the models are evaluated using the three types of matching. 'None' corresponds to using no external knowledge. Adding external knowledge improves the performance of the models in three out of four cases.

Model	EX-MATCH			SEM-MATCH		SEM/EX-MATCH	
	Precision	Recall	Accuracy	BLEU	BLEURT	BLEU	BLEURT
Ponkiya et al. (2020)	0.0	0.0	60.8	23.1	44.9	13.8	26.8
Rand Few-Shot (5)	37.3	11.0	55.3	27.7	40.2	18.5	25.1
Rand Few-Shot (10)	62.1	21.4	58.2	27.6	39.3	22.3	28.2
KNN Few-Shot (5)	68.7	43.6	69.1	**29.9**	46.1	33.1	41.4
KNN Few-Shot (10)	**67.1**	**50.5**	**69.9**	**29.9**	**46.9**	**35.2**	**43.7**

Table 6: Performance of T5 model without any finetuning. Ponkiya et al. (2020) corresponds to the zero-shot setting adapted from the corresponding paper. Few-shot techniques use either five or ten example demonstrations. In 'Rand' the few-shot examples are chosen randomly while in 'KNN' the nearest neighbours of the query are chosen as the few-shot examples. Availability of annotated examples from PRONCI helps to substantially improve the performance of the model. Overall performance remains inferior to the finetuned models.

7 Experiments

In this section, we address the following questions:

1. How do UNIGEN and MTGEN compare with each other and what benefit does adding external knowledge provide to these models?

2. What is the performance difference between few-shot learning and supervised training?

3. How do individual components of the noun compound influence the model predictions?

7.1 Performance of Supervised Models

In Table 5, we show the results of both, the multi-task model, MTGEN and the unified generation model, UNIGEN (Section 5).

We find that the UNIGEN model outperforms the MTGEN model in overall performance but leads to a modest drop in the compositionality classification performance. For example, in the case where no additional knowledge is used, UNIGEN leads to a higher SEM/EX-MATCH score with both BLEU and BLEURT leading to an increase of (2.4, 1.1) pts. But UNIGEN achieves a lower classification score with the EX-MATCH accuracy reducing by 0.8%. We attribute this observation to the fact that MTGEN uses a separate module that enables it to be tuned better for the classification task. However, UNIGEN performs better in overall performance as both the encoder and decoder can benefit from positive transfer between the tasks.

By adding knowledge to the model, using the prompting described in Table 3, at both training and testing time, we see gains in performance in three out of four types of knowledge added. Using information of the proper noun from Wikipedia often reduces the performance due to incorrect entity linking. Among the the remaining three sources of knowledge, we find that WordNet-NN leads to the maximum increase in performance in three of the four settings. We find that the predicted interpretations are often biased to re-use words that occur in the knowledge prompts and this leads to higher scores in case of less frequently occurring compounds. For instance, the prediction changes from "Kirati community is a group of Kirati" to "Kirati community are people of Kirati", when added with

the knowledge, "Major groups of Kirati community follows Buddhism". Using student paired t-test we find that improvements are statistically significant with p-value of $3.78e^{-10}$ of BLEURT scores averaged over all 5 seeds. We do not find additional improvements when multiple knowledge sources are added simultaneously (Appendix E).

Predictions of UNIGEN trained with sentence knowledge are rated to be 72% correct when checked manually on a sample of 100 sentences. This indicates a significant scope for improvement, when compared to the upper bound of 93% data quality (Section 4).

We conduct two further experiments on the trained UNIGEN model to understand the strength of semantic matching used and the effectiveness of the model on the related task of common noun compound interpretation.

Template scoring: To test the effect of template word matching on BLEU and BLEURT scores, we introduce a dummy relation: i.e., the prediction for every non-compositional example is forced to be 'noun-compound is none of common-noun'. This ensures that only template words match, but the semantic meaning is wrong. On re-computing the SEM-MATCH scores of UNIGEN, this reduces the BLEU score from 34.1 to 22.9 and BLEURT score from 46.7 to -3. This follows the expected trend as BLEU gives partial scores to template matches, but BLEURT focuses on the overall semantic meaning.

SemEval evaluation: When UNIGEN is evaluated on the free noun compound paraphrasing task of SemEval 2013 Task 4 (Hendrickx et al., 2013), it achieves an isomorphic score of 72.8 compared to 80.1 reported by Ponkiya et al. (2018). We attribute this to different interpretation styles with PRONCI focusing on detailed relations (average length of 6.9 words) compared to SemEval (average length of 5.1 words), leading to slightly lower scores with word match heuristics adopted by the task.

7.2 Performance of few-shot learning

State-of-art models for free paraphrase interpretations of *common* noun compounds (Ponkiya et al., 2020) uses the zero-shot generation capabilities of T5 and show that they outperform supervised models. To check if the same holds for the PRONCI dataset, we experiment with zero-shot generation. Similar to (Ponkiya et al., 2020), we use the masked template, "$w_1 w_2$ is a <*extra_id_0*> the w_1", where T5 fills in the missing words in place of <*extra_id_0*>.

We further experiment with few-shot learning, where K training examples are chosen as part of the prompt which the model can use to perform in-context learning and generate the prediction for the given input. No additional knowledge is used in these set of experiments. These K examples can either be chosen randomly or the nearest neighbours to the input query can be chosen, where cosine distance between the input and a training example is computed after embedding them with a pre-trained T5-Encoder (Liu et al., 2022a). We experiment with K = 5 or 10. The limitations of context size in the pretrained models prevent us from testing with higher values of K.

In Table 6, we find that the zero-shot performance trails behind the best few-shot model with a decrease of 21.4, 41 pts in BLEU/EX, BLEURT/EX, respectively. This is partly because of the variety of examples in the PRONCI dataset, which cannot be fit into specific templates and the inability of zero-shot model to handle non-compositional examples. In few-shot learning, expanding the prompt size and dynamically choosing the prompt examples helps achieve higher performance but the performance still remains lower than the fully-supervised UNIGEN model which is still 11.2, 15.3 pts higher in BLEU/EX, BLEURT/EX.

7.3 Proper noun vs. Common noun

The interpretation of a proper noun compound depends on both the proper noun and common noun present in it. To study how each of the two nouns influence the prediction, we randomly shuffle the their characters in both input and gold annotation.

In Table 7, we find that common noun has a larger effect on the model performance as shuffling its characters leads to a significant drop performance of (5.8, 17.6, 35) pts in (BLEU/EX, BLEURT/EX, EX-MATCH Accuracy%). Comparatively, the proper noun results in a much smaller drop of (3.1, 8, 16.3) pts in the three evaluation metrics.

8 Application to Open IE

To demonstrate the downstream value of the noun compound interpretations, we add them to a state-of-art Open IE system, Gen2OIE (Kolluru et al., 2022), and generate new extractions that capture implicit relations. We apply this on a corpus of

Shuffle	EX-MATCH	SEM/EX-MATCH	
	Accuracy	BLEU	BLEURT
None	**78.7** ± 1.40	**46.7** ± 1.12	**58.6** ± 0.94
NNP	62.4 ± 0.97	43.6 ± 1.01	50.6 ± 0.44
NN	43.7 ± 1.02	40.9 ± 0.15	41.0 ± 0.16

Table 7: UNIGEN evaluated after random shuffling of characters in the proper (NNP) or common (NN) noun.

21,228 COVID-19 news headlines that contain proper noun compounds like COVID-19 outbreak, Rohingya refugee, etc (Aslam et al., 2020).

Integration: To achieve this, we train a Seq2Seq model that takes as input the sentence concatenated with the interpretation of the PNC present in it and outputs an interpretation-augmented sentence. For example, the sentence, "Workers sound alarm on *Covid-19 outbreak*" and the interpretation, "*Covid-19 outbreak* is an outbreak of Covid-19" are integrated to get the following output, "Workers sound alarm on outbreak of Covid-19". Considering the simplicity of the task, we annotate a small set of 200 examples of this kind and use it to train a Seq2Seq model. Since this style of integration converts the implicit relation in the noun compound to an explicit form, it allows for the Open IE system to add new relations that were missing earlier.

Processing: We experiment with a high precision rule that post-processes an extraction to generate a new one, whenever the extraction contains a PNC at the start of its object. For example, if the original extraction is (Workers; sound alarm on; COVID-19 outbreak), and the corresponding integrated extraction is (Workers; sound alarm on; outbreak of COVID-19), then the rule generates a new extraction by moving words till the proper noun back into the relation. In this case, we get the extraction, (Workers; sound alarm on outbreak of; COVID-19) – thus exposing a direct relationship between workers and COVID-19, which was not present earlier. The overall pipeline is shown in Figure 3.

We find that extractions generated using this pipeline leads to an increase in yield of 7.5% where the added extractions have a precision of 85%, compared to a precision of 82.2% of the original extractions, as determined on a random sample of 500 extractions. We note that the method can use any Open IE system without any additional finetuning to produce the noun-compound extractions.

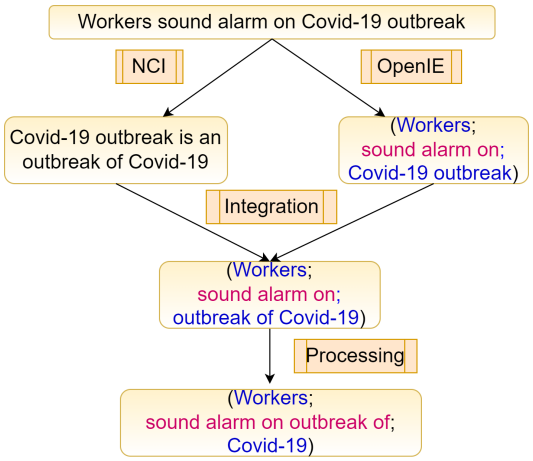

Figure 3: Open IE Pipeline. Postprocessing of the extraction integrated with noun compound interpretation, generates the new extraction.

9 Conclusion

In this work we develop the novel task of semantic intepretation of proper noun compounds. We present the PRONCI dataset for the task and test performance of various neural models. We show the downstream utility of generated interpretations by integrating it with an Open IE system that results in generation of new extractions involving implicit relations. Linguistic characterization of non-compositionality and utilizing additional sources of knowledge present scope for further work.

10 Limitations

The proposed models are evaluated on a specific test set, which may not be representative of all the types of examples that it may encounter during deployment. Due to the use of pretrained models, the system may also exhibit possible biases that have been discovered in the pretrained models.

Acknowledgements

Keshav is supported by a TCS Fellowship. Mausam is supported by grants from Huawei, Google, Verisk, and a Jai Gupta Chair Fellowship. We thank KnowDis team for their help in data annotations and HPC, IIT Delhi for the computational resources. This work was supported in part by a research grant no. 2088 from the Israeli Ministry of Science and Technology.

References

Kisuh Ahn, Johan Bos, David Kor, Malvina Nissim, Bonnie L Webber, and James R Curran. 2005. Question answering with qed at trec 2005. In *TREC*.

Artemis Alexiadou. 2019. Proper name compounds: a comparative perspective. *English Language & Linguistics*, 23(4).

Faheem Aslam, Tahir Mumtaz Awan, Jabir Hussain Syed, Aisha Kashif, and Mahwish Parveen. 2020. Sentiments and emotions evoked by news headlines of coronavirus disease (covid-19) outbreak. *Humanities and Social Sciences Communications*, 7(1).

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by machine of complex nominals: Getting it right. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*.

Renu Balyan and Niladri Chatterjee. 2015. Translating noun compounds using semantic relations. *Computer Speech & Language*, 32(1).

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *International Joint Conference on Artificial Intelligence (IJCAI), 2007*.

Tine Breban, Tine Breban, and Julia Kolkmann. 2019. Different perspectives on proper noun modifiers. *English Language & Linguistics*, 23(4).

Samuel Broscheit, Kiril Gashteovski, Yanjie Wang, and Rainer Gemulla. 2020. Can we predict new facts with open knowledge graph embeddings? a benchmark for open link prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2009. SemEval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2011. An analysis of open information extraction based on semantic role labeling. In *Proceedings of the sixth international conference on Knowledge capture*.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Murhaf Fares. 2016. A dataset for joint noun-noun compound bracketing and interpretation. In *Proceedings of the ACL 2016 Student Research Workshop*.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*.

Swapnil Gupta, Sreyash Kenkre, and Partha Talukdar. 2019. CaRe: Open knowledge graph embeddings. 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)*.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*.

Matthew Honnibal, Ines Montani, Matthew Honnibal, Henning Peters, Sofie Van Landeghem, Maxim Samsonov, Jim Geovedi, Jim Regan, György Orosz, Søren Lind Kristiansen, Paul O'Leary McCann, Duygu Altinok, Roman, Grégory Howard, Sam Bozek, Explosion Bot, Mark Amery, Wannaphong Phatthiyaphaibun, Leif Uwe Vogelsang, Björn Böing, Pradeep Kumar Tippa, jeannefukumaru, GregDubbin, Vadim Mazaev, Ramanan Balakrishnan, Jens Dahl Møllerhøj, wbwseeker, Magnus Burton, thomasO, and Avadh Patel. 2019. explosion/spaCy: v2.1.7. In *explosion/spaCy: v2.1.7*.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020a. OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam, and Soumen Chakrabarti. 2020b. IMoJIE: Iterative memory-based joint open information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Keshav Kolluru, Mohammed Muqeeth, Shubham Mittal, Soumen Chakrabarti, and Mausam. 2022. Alignment-Augmented Consistent Translation for Multilingual Open Information Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Mark Lauer. 1995. Designing statistical language learners: Experiments on noun compounds. In *Ph.D.thesis, Ph. D. thesis, Macquarie University*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022a. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*.

Qi Liu, Dani Yogatama, and Phil Blunsom. 2022b. Relational memory augmented language models. *arXiv preprint arXiv:2201.09680*.

Pratyush Maini, Keshav Kolluru, Danish Pruthi, and Mausam. 2020. Why and when should you pool? analyzing pooling in recurrent architectures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Mausam. 2016. Open information extraction systems and downstream applications. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*.

Preslav Nakov. 2013. On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, 19(3).

Harinder Pal and Mausam. 2016. Demonyms and compound relational nouns in nominal open IE. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*.

Soma Paul, Prashant Mathur, and Sushant Kishore. 2010. Syntactic construct : An aid for translating English nominal compound into Hindi. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*.

Girishkumar Ponkiya, Rudra Murthy, Pushpak Bhattacharyya, and Girish Palshikar. 2020. Looking inside noun compounds: Unsupervised prepositional and free paraphrasing using language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*.

Girishkumar Ponkiya, Kevin Patel, Pushpak Bhattacharyya, and Girish K Palshikar. 2018. Towards a standardized dataset for noun compound interpretation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Anette Rosenbach. 2007. Emerging variation: Determiner genitives and noun modifiers in english. *English Language & Linguistics*, 11(1).

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

Vered Shwartz and Chris Waterson. 2018. Olive oil is made *of* olives, baby oil is made *for* babies: Interpreting noun compounds using paraphrases in a neural model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

Stephen Soderland, L. Kim, and Natalie Hawkins. 2015. A language model for extracting implicit relations. In *https://www.cs.rochester.edu/u/gkim21/papers/IMPLIE-2015.pdf*.

Stephen Tratz. 2011. *Semantically-enriched parsing for natural language understanding*. University of Southern California.

Shikhar Vashishth, Prince Jain, and Partha P. Talukdar. 2018. CESI: canonicalizing open knowledge bases using embeddings and side information. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*.

Pat Verga, Haitian Sun, Livio Baldini Soares, and William Cohen. 2021. Adaptable and interpretable neural MemoryOver symbolic knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

"*Covid vaccine* is against Covid but *Oxford vaccine* is made at Oxford!" Semantic Interpretation of Proper Noun Compounds (Appendix)

A Quality Assesment of Evaluation Metrics

For evaluating the quality of the metrics that are used for evaluating the model predictions, in particular, the semantic matching component Section 6, we manually annotate the quality of model predictions with respect to gold using a 3-index scale. The scale indicates whether the quality of the prediction is bad, average or good. This is done only for the cases where the gold annotation indicates that the compound is compositional and the prediction of the model is also a paraphrase, as semantic matching is applicable only in these cases. A total of 1500 examples are annotated out of which 500 are used for finetuning the learned metrics such as BLEURT. On the remaining 1K examples, we compute the Pearson and Kendall correlation between the scores assigned by the evaluation metric and the human annotated scores. We report the results in Table 8 for five evaluation metrics which include BLEU, BLEURT with and without finetuning on both the base and large variants. We find that the fine-tuned BLEURT outperforms both BLEU and the un-trained BLEURT. It specifically outperforms BLEU by a significant margin from 0.28 to 0.57 in Pearson correlation and 0.23 to 0.46 in Kendall correlation. We find that the performance of both the base and large variants of BLEURT perform similarly after being finetuned and a minor difference exists in their untuned variants.

We note that the correlation of 0.57 is on par with the current state of NLG metrics. For example, Chen et al. (2020), reports a correlation of 0.45-0.60 for standard metrics such as BLEU, BERTScore (Zhang et al., 2020) on short-text evaluation. To further encourage research in building better generation metrics, we release the human judgements of the interpretations.

Metric	Pearson $\lvert\rho\rvert$	Kendall τ
BLEU	0.28	0.23
BLEURT-base	0.43	0.37
BLEURT-large	0.49	0.4
BLEURT-base *(tuned)*	0.56	**0.46**
BLEURT-large *(tuned)*	**0.57**	**0.46**

Table 8: Quality of metrics evaluated using Pearson and Kendall rank correlation. (tuned) indicates models that are fine-tuned on 500 manually evaluated comparisons.

B Random Split of PRONCI

In this section, we evaluate the results of UNIGEN and MTGEN on a random split of the PRONCI dataset, where the 22,500 examples are randomly split into 17,500 training, 2,500 validation and 2,500 testing examples. The results are reported in Table 9 and Table 10. We find that the performance is higher compared when split according to common nouns. This can be attributed to the lack of intersecting common nouns between the training and evaluation sets that could have provided additional clues. This leads to a drop in (BLEU/EX, BLEURT/EX, EX Acc%) scores of (5.7, 5.4, 3.3) pts in MTGEN and (5.3, 4.6, 2.9) pts in UNIGEN.

C Effect of Pretraining

To understand the effect pretraining has on the effect of model performance for the task of semantic interpretation of proper noun compounds, we re-train the UNIGEN on the NOUN split starting from random initialization, instead of using T5-base, the default in all of our experiments. We also experiment with using T5-large. We report the results in Table 11. We find that Random initialization is considerably worse, where the scores reduces from 46.7 to 33.9 in BLEU/EM and 58.6 to 30.5 in BLEURT/EM. This indicates that pretrained initialization plays a significant role in the final performance on the task. Moreover, on experimenting with the larger model, T5-large, we find a slight increase in scores from (46.7, 58.6, 78.7) to (47.7, 58.7, 79.4) in (BLEU/EM, BLEURT/EM, CMP). Thus the task can benefit from scaling of the language models as they typically gain more information about the common and proper nouns.

D Error Analysis

We analyze the mistakes made by the UNIGEN model trained with Sentence Knowledge to find potential scopes for improvement. We divide them into the following categories -

1. Lack of word sense disambiguation: We notice mistakes in the model predictions in cases when some words have multiple meanings.

Model	Knowledge	EX-MATCH			SEM-MATCH		SEM/EX-MATCH	
		Precision	Recall	Accuracy	BLEU	BLEURT	BLEU	BLEURT
MTGEN	None	**78.2** ± 1.14	74.5 ± 1.48	82.8 ± 0.25	40.5 ± 0.58	63.8 ± 0.36	50.0 ± 0.39	62.9 ± 0.29
	Sentence	76.9 ± 1.70	78.6 ± 1.38	83.3 ± 0.69	40.4 ± 0.43	63.2 ± 0.30	51.1 ± 0.25	63.4 ± 0.50
	WordNet-NN	76.4 ± 1.30	80.5 ± 1.95	**83.5** ± 0.36	40.8 ± 0.63	63.3 ± 0.40	51.8 ± 0.55	63.8 ± 0.47
	Wiki-NNP	51.7 ± 1.04	94.7 ± 0.82	65.2 ± 1.41	25.9 ± 1.26	36.0 ± 3.80	42.9 ± 0.44	46.0 ± 1.14
	NER-NNP	75.4 ± 2.19	80.5 ± 3.06	82.9 ± 0.45	40.5 ± 0.79	63.4 ± 0.62	51.4 ± 0.29	63.5 ± 0.26
UNIGEN	None	71.7 ± 0.68	83.4 ± 1.07	81.6 ± 0.21	41.5 ± 0.16	63.7 ± 0.17	52.0 ± 0.24	63.2 ± 0.15
	Sentence	72.1 ± 0.32	83.6 ± 0.44	81.9 ± 0.19	41.3 ± 0.19	63.4 ± 0.45	52.0 ± 0.12	63.3 ± 0.17
	WordNet-NN	71.0 ± 1.71	**86.2** ± 1.23	81.7 ± 0.88	**42.0** ± 0.40	64.0 ± 0.39	**52.9** ± 0.34	**63.8** ± 0.42
	Wiki-NNP	68.6 ± 2.08	68.2 ± 1.93	76.5 ± 0.82	26.1 ± 0.78	39.0 ± 2.18	38.7 ± 0.42	45.3 ± 1.25
	NER-NNP	71.9 ± 0.98	81.8 ± 1.70	81.3 ± 0.25	41.6 ± 0.34	**64.2** ± 0.68	51.6 ± 0.42	63.1 ± 0.49

Table 9: Performance of the two models, MTGEN and UNIGEN on the randomly split PRONCI dataset trained under five different knowledge settings.

Model	EX-MATCH			SEM-MATCH		SEM/EX-MATCH	
	Precision	Recall	Accuracy	BLEU	BLEURT	BLEU	BLEURT
Ponkiya et al. (2020)	0.0	0.0	62.8	22.9	44.1	14.4	27.7
Rand Few-Shot (5)	53.7	1.2	63.0	27.6	41.2	17.7	26.2
Rand Few-Shot (10)	37.7	33.6	54.4	28.8	42.2	24.7	29.7
KNN Few-Shot (5)	**70.5**	53.0	74.3	34.8	51.7	38.7	48.0
KNN Few-Shot (10)	68.4	**60.1**	**74.8**	**35.4**	**53.2**	**41.0**	**50.3**

Table 10: Performance of T5 model without any finetuning on the random split of PRONCI dataset.

Init	EX-MATCH	SEM/EX-MATCH	
	Accuracy	BLEU	BLEURT
Random	63.9 ± 1.98	33.9 ± 1.25	30.5 ± 1.27
T5-base	78.7 ± 1.40	46.7 ± 1.12	58.6 ± 0.94
T5-large	**79.4** ± 0.11	**47.7** ± 0.29	**58.7** ± 0.35

Table 11: Performance of the UNIGEN model on the PRONCI dataset trained using different initializations of the Seq2Seq model. Random initialization leads to huge drop in performance.

Knowledge	EX-MATCH	SEM/EX-MATCH	
	Accuracy	BLEU	BLEURT
Sentence	79.3 ± 0.55	**47.9** ± 0.41	**59.5** ± 0.57
+WNet-NN	77.4 ± 2.14	46.5 ± 1.48	57.4 ± 1.63
+Wiki-NNP	74.0 ± 1.62	38.9 ± 4.21	46.1 ± 6.38
+NER-NNP	**79.4** ± 0.23	47.0 ± 0.52	58.9 ± 0.40

Table 12: Performance of the UNIGEN model on PRONCI dataset trained with additional sources of knowledge added over Sentence knowledge. The additional sources do not provide further benefits.

The model defaults to choosing the one with most frequent usage and not disambiguating properly based on the context. For example, the the interpretation, "*Sunday strip* is a comic printed on a Sunday" is mistaken as "*Sunday strip* is a show on Sunday", even when the sentence contains sufficient clues for the same. The given sentence is "In a few cases, the topper introduced characters later developed into a successful Sunday strip."

2. Non Informative predictions: Although predictions are not wrong they are often not very informative. For example, the model produces the following interpretation, "*EU economies* are based in EU" compared to the more detailed gold "*EU economies* are the financial condition of EU members".

3. Errors in evaluation and mistakes in Gold: In some cases, the evaluation metric is unable to capture semantic similarity. For example, the model prediction "Baltimore hospitals are located in Baltimore" and the gold, "Baltimore hospitals are medical institutions in Baltimore", has a BLEURT score of only -0.11.

E Adding multiple sources of knowledge

In Section 7, we observed statistically significant benefits to model performance after adding information about the noun compound from various sources of knowledge. We also experiment with adding information from multiple source of knowledge to see if it can further augment the model performance. On taking the best performing Sen-

tence knowledge in the UNIGEN model on NOUN split, we add the remaining three sources of knowledge and report their performance in Table 12. We find that it results in a slightly decrease in performance in case of WNet-NN and NER-NNP and in case of Wiki-NNP the decrease is much greater because of the reduced quality of Wikipedia entities. We attribute this to possible confusion arising from disparate sources of knowledge that highlight different parts of the noun compound.