ANA* at SemEval-2020 Task 4: mUlti-task learNIng for cOmmonsense reasoNing (UNION)

Anandh Perumal, Chenyang Huang, Amine Trabelsi, Osmar R. Zaïane
Alberta Machine Intelligence Institute, University of Alberta
{anandhpe, chenyangh, atrabels,zaiane}@ualberta.ca

Abstract

In this paper, we describe our mUlti-task learNIng for cOmmonsense reasoNing (UNION) system submitted for Task C of the SemEval2020 Task 4, which is to generate a reason explaining why a given false statement is non-sensical. However, we found in the early experiments that simple adaptations such as fine-tuning GPT2 often yield dull and non-informative generations (e.g. simple negations). In order to generate more meaningful explanations, we propose UNION, a unified end-to-end framework, to utilize several existing commonsense datasets so that it allows a model to learn more dynamics under the scope of commonsense reasoning. In order to perform model selection efficiently, accurately and promptly, we also propose a couple of auxiliary automatic evaluation metrics so that we can extensively compare the models from different perspectives. Our submitted system not only results in a good performance in the proposed metrics but also outperforms its competitors with the highest achieved score of 2.10 for human evaluation while remaining a BLEU score of 15.7. Our code is made publicly availabled [1].

1 Introduction

Common sense reasoning is one of the long-standing problems in natural language understanding. Previous work on modeling common sense knowledge deals mainly with indirect tasks such as co-reference resolution (Sakaguchi et al., 2019), or selecting the plausible situation based on the given subject or scenario (Zellers et al., 2018).

In this paper, we present our system that we devised to tackle Task C, Explanation (Generation), of the SemEval 2020 Task 4 - Commonsense Validation and Explanation (ComVE). Given a false or non-sensical statement, the task consists of generating the reason why the given statement does not make sense. We propose a mUlti-task learNIng for cOmmonsense reasoNing (UNION). It combines datasets including ComVE (Wang et al., 2019), OpenBook (Mihaylov et al., 2018), Common sense Explanation (CoS-E) (Rajani et al., 2019) and Open Mind Common Sense (OMCS) (Singh et al., 2002) in a multi-task framework. The backbone of UNION is a large pre-trained language model – GPT2.

We compare the proposed system to different baselines and report a significant improvement in BLEU score and other evaluation metrics. Our proposed model achieves a human evaluation score of 2.10, which ranked first on the final leader board for Task C of SemEval 2020 Task 4. In our initial submission, we used ComVE, CoS-E, and OpenBook datasets for training. In that case, a BLEU score of 15.7 is achieved. Pretraining the model with OMCS dataset further improved the BLEU score by 0.7. In addition, we show some of our generations in the appendix.

2 Background and Related Work

The common sense reasoning in machines has been one of the most challenging problems, and a major critical missing component of Artificial Intelligence (AI). It could be helpful in various aspects of day-to-day life (Gunning, 2018). For any AI system to generate commonsense reasoning, it needs to understand

*https://www.amii.ca/ana-automated-nursing-agent/

[1]https://github.com/anandhperumal/ANA-at-SemEval-2020-Task-4-UNION

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 367–373
Barcelona, Spain (Online), December 12, 2020.

and build a representation of the given situation (Mueller, 2014). For instance, let us consider the following statement "You will never find a dog that likes to eat meat." Humans may have intuitively built a knowledge graph containing the facts that "dogs are carnivorous" and "a carnivorous is a meat-eater". Thus, based on it, they may conclude that the first statement is a false statement. Similarly, for a machine to answer questions about general situations or facts, a knowledge graph or formal logic system could be beneficial. Early work by John McCarthy (1960) proposed a system that uses formal logic for commonsense reasoning. Thus, the system can induce commonsense reasoning given all the possible axioms about the studied domain or the world. However, it is not feasible to generate all the reasonable axioms about the world. This gave rise to several other logic-based approaches for commonsense reasoning and numerous work with the aim of creating huge logic-based ontologies, e.g., situation calculus (Fikes and Nilsson, 1971), YAGO (Suchanek et al., 2007), DBpedia (Auer et al., 2007), Event2mind (Rashkin et al., 2018). However, commonsense reasoning over a particular knowledge often acts as a lookup table as it lacks sufficient semantics to formulate a complete sentence as a reason. Other models (Rajani et al., 2019; Devlin et al., 2018; Kwiatkowski et al., 2019; Taylor, 1953) have achieved competitive performance in question answering tasks given a comprehension or a document as a source of information. All these models help to simulate common sense reasoning in the language model. However, they still rely on a passage as a source of information, which makes them limited to specific domains.

An AI system which processes common sense like humans, in the context of question answering or dialogue generation, should not depend on a source of information each time it produces a response to a query. Instead, it should learn and generate a response from the previously learned information. Thus, our proposed model does not rely on any other source of information during inference other than its prior learned knowledge. We train UNION on multiple commonsense datasets ComVE, CoS-E, and OMCS.

3 Proposed systems

The language generation task is to model the conditional probability $P(Y|X)$, where X and Y are both sequences of words. In this work, we use the 36-layer decoder-only transformers based on the architecture of the Generative Pre-trained Transformer (GPT2) (Radford et al., 2019) as the backbone of the language generation task for all the presented models. The decoder only transformer architecture is similar to the original decoder transformer architecture proposed by Vaswani et al. (2017). The main difference is that in the decoder-only transformer case, the decoder component does not have a multi-head attention for the encoder input. All our models are initialized with the large weights of the pre-trained language model GPT2. In this section, we review the baseline models and the architecture of our proposed UNION model.

3.1 Baseline models

Language generation baseline: For Task C Explanation (Generation), the model needs to generate the reason why a given false statement is against common sense. The data is given in the format of $\{X^i, \mathbf{Y}^i\}$, where each statement X^i is paired with a three possible explanations $\mathbf{Y}^i = (Y_1^i, Y_2^i, Y_3^i)$.

We formulate Task C as a sequence-to-sequence generation task. A false statement X^i is considered as the source to the language model while one of the explanations from $\mathbf{Y}^i$ becomes the target. Hence, as a first step, we transform our dataset format. It is no longer of the format $(X^i, Y_1^i, Y_2^i, Y_3^i)$, but of the format (X^i, Y_j^i), where $Y_j^i \in \mathbf{Y}^i$ for $j = 1, 2, 3$. As a second step , we train our baseline language model GPT2 with the new dataset format. The final goal is to estimate the conditional probability distribution $P(Y_j^i|X^i)$ of the target statement Y_j^i given the source statement X^i, where $j = 1, 2, 3$.

Multi-task learning baseline: We train a language model in the Multi-task learning (MTL) setting. The model has two heads, one for language model and another one for the classification task. Caruana (1997) has briefly expressed the advantage of MTL: "*MTL improves generalization by leveraging the domain-specific information contained in the training signals of related tasks*". Moreover, Raffel et al. (2019), Liu et al. (2019), and Devlin et al. (2018) have shown that training language models using an MTL technique helps the language model to learn and generalize better.

We chose to train an MTL model using ComVE's Task B and Task C datasets because of their similarity. Task B of ComVE is a multi-choice classification problem. For each false statement X^i, there exist three

plausible explanations from $\hat{\mathbf{Y}}^{\mathbf{i}} = (\hat{Y}_1^i, \hat{Y}_2^i, \hat{Y}_3^i)$. Although, only one of these is correct, they all have the same syntactic structure and similar wording, and only differ by few words. Moreover, all of the correct explanations present in Task B are a subset of explanations $\mathbf{Y}$ provided in Task C (Generation). Thus, we hypothesized that adding this task in an MTL setting would help the model better discriminate between a valid and an invalid explanation, and consequently learn to recognize the subtle differences allowing it to generate better valid reasons. Moreover, training on more similar examples to those in Task C would make the generated text more aligned with the syntactic structure and keywords of the task's data.

We convert the dataset format from a multi-label to a binary classification format $(X^i, \hat{Y}_k^i, b)$, where b is a binary label (correct or not), and $k = 1, 2, 3$. The Task B and C have the same set of false statements X as input. Thus, we pair the dataset according to the false statement. The new data format is the following: $\{(X^i, Y_j^i), (X^i, \hat{Y}_k^i, b)\}$. We train the language modeling and the classification heads in parallel. The loss is computed by taking the summation of both heads' losses.

3.2 Why UNION?

A language model is a probability distribution learned over the sequence of tokens in the training corpus. Thus, the performance of the model strongly depends on the quality of the data used for its training. When we investigated the explanations generated by the baseline and MTL models, it seemed that these tended to often negate the given input statement. Although this may still make them valid explanations, it did not necessarily make them express reasoned explanations on why the input is a false statement. Part A of Table 3 provides examples of two valid explanations where the first is just a negation of the input whereas the second is a better systematic explanation of why the given statement is false. The root cause for generating a simple negation of a statement over a reasoned explanation is the ComVE dataset itself. More than one-third of it contains negating statements, which most probably signals the model to learn to negate any given input. To deal with this issue, we may either remove the explanations that are negation to the false statement or increase the dataset by adding better explanations to it. Since the dataset for explanation generation is limited, deleting any explanations may have a negative impact, while creating a new dataset is a tedious task in itself. Therefore, we resorted to using other commonsense related datasets, CoS-E, OpenBook, and OMCS , along with the ComVE dataset, to train the language model to generate better responses. These additional datasets treat different issues. OpenBook is a question answering dataset related to science facts. CoS-E is for general commonsense question answering. OMCS contains common knowledge facts. Training all four together leads to two main difficulties or questions to answer:

1. Each dataset is related to a different issue while our main task is to generate an explanation for a false statement. How can we force the model to generate an explanation response that is specific to the false statement and not any random generic statement related to it?
2. Each dataset has a different number of classification choices. How do we train all of them together in an MTL setting?

Our proposed architecture, described in Section 3.3, solves the first problem with the help of a contextual keyword and the second problem by merging all the classes.

3.3 UNION

The backbone of the mUlti-task learNIng for cOmmonsense reasoNing (UNION) architecture in Figure 1 is a decoder-only transformer, as described above. The UNION model is categorized into four major layers, shared, pre-trained, semi-shared, and task-specific layer. The first layer is a shared layer among all the other following layers. The second layer is a language modeling head. It is used to pre-train the UNION model with the OMCS dataset before training with ComVE, CoS-E, and OpenBook. The OMCS dataset has over a million statements as facts and common knowledge, e.g., "You are likely to find a shelf in a cupboard.". The pre-training helps the model to learn about the general facts and knowledge about the world. The third layer is semi-shared language modeling (SSLM) head shared among data from ComVE, CoS-E, OpenBook. It is used to train the explanation generation for all three datasets.

We use a contextual keyword (c) to solve the first issue described in Section 3.2, i.e., to generate a response or an explanation that is specific to a particular dataset. The contextual keyword is a unique token for each dataset. During training, we condition the Y on (c, X). Therefore, we no longer estimate

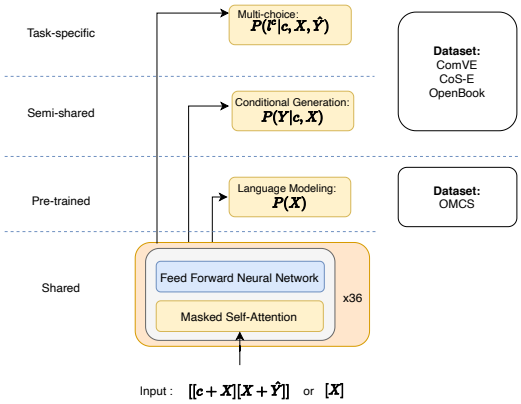

Figure 1: UNION Architecture

False Statement	We use book to know the time
Referential Reasons	a) A book is used to study b) A book does not have the ability to show what time it is. c) Books don't tell the time
Generated Explanation	Book is not a timekeeping device.

Table 1: UNION model generated explanation

the $P(Y|X)$, but we estimate the $P(Y|c, X)$, while during inference, we condition the false statement of the ComVE dataset to the particular contextual keyword to generate an explanation. We call this layer semi-shared because we share the same language modeling head layer for ComVE, CoS-E, and OpenBook while we also condition the generation for each given task based on the contextual keyword.

Finally, we have a task-specific layer. The task-specific layer is used for multi-class classification (MC), and the same contextual keyword for the SSLM head layer is used in the MC layer too. The final classification layer size is twelve (first three classes for ComVE, four for OpenBook, and five for CoS-E). Combining all the classes helps to mitigate the second problem described in the previous section of the difference in the number of classes between datasets.

Therefore, during training and inference, with the help of the contextual keyword, we ignore the labels which are not relevant for the particular dataset by assigning zero and normalizing the probability distribution over the remaining relevant labels. For example, during the training of the ComVE dataset using the contextual keyword, we are assigning zero probability to the labels from four to twelve (i.e., those related to OpenBook and CoS-E), and the probability distribution is normalized over the first three labels. So, we estimate the $P(l^c|c, X, \hat{Y})$, where $\hat{Y}$ is a concatenation of all the potential explanations for X, l^c is the label, and c is the contextual keyword.

4 Experiments

4.1 Evaluation metrics

The BLEU score (Papineni et al., 2002) is widely used in tasks such as machine translation (Sutskever et al., 2014; Chiang, 2005). It calculates the overlapping between the candidates and the reference text. However, similar to dialogue generation, our experimental results show that the BLEU score is not an ideal measurement since it does not tolerate diversity. For instance, as shown in Table 1, the generated explanation example "Book is not a timekeeping device." has little overlap with the reference but it should still be considered as a good generation. As a remedy, the submitted systems are evaluated by human evaluations. The BLEU score and the human evaluation are carried out by the task organizers, where the human evaluation score is based on the agreement between three reviewers.

In order to better evaluate the performance of the proposed systems and conduct ablation studies, we use several auxiliary metrics to assess the quality of the explanations generated by various models.

Perplexity: We first use perplexity to measure the [grammatical/syntactical] correctness of the generated text. Particularly, we suggest two variants of perplexity: the general perplexity (*ppl-gen*) and the target corpus perplexity (*ppl-trg*). We measure *ppl-gen* by using the GPT2 language model (LM) head directly, as it is pre-trained on large scale corpus (Reddit, Wiki). It gives an indication on the fluency of the generated content in general. In the meantime, we also want to assess how well the generations can be in terms of fitting into the dialect of the target corpus. We achieve this by training an n-gram language model

Models	BLEU	PPL - Gen.	PPL - Trg.	EA	UNI	Length
Baseline	10.36	970.05	495.35	0.86I	3.55 ± 1.77	5.5 ± 1.97
Baseline + MTL	12.4	357.59	331.22	0.93	3.31 ± 1.68	5.59 ± 1.89
UNION w/o CoSE	13.28	62.89	238.64	0.96	5.82 ± 2.10	8.51 ± 2.08
UNION w/o OpenBook	13.75	142.19	260.38	0.95	4.29 ± 1.87	6.46 ± 2.19
UNION w/o OMCS	15.7	194.66	243.83	0.94	4.29 ± 1.79	6.41 ± 2.06
UNION	16.36	135.1	212.1	0.97	4.53 ± 2.05	6.59 ± 2.3

Table 2: Ablation study results on different proposed models

with Kneser-Ney smoothing[2]. The n-gram language model is trained based on Task B and C datasets.

Informative: As shown in Table 3, the generations produced by the baseline models are often derived from the false statement X^i by changing very few words. To measure the impact of the change in the generated responses, we propose two auxiliary metrics: *Estimated Approval (EA)* and *Uniqueness (UNI)*.

False Statement	Explanation
A) The chocolate cried.	1) Chocolate doesn't cry.
	2) Chocolate is an inanimate, non human thing and cannot cry.
B) Sugar is used to make coffee sour.	1) Sugar is used to make coffee sweet. (UNION)
	2) Sugar is not used to make coffee sour. (Baseline)

Table 3: Examples from ComVE explanation (Generation), and UNION and Baseline models

The UNI and Length metric measures the added amount of information of an answer. UNI calculates the number of tokens that are not present in the given input and Length measure the length of each explanation. A high UNI and Length value suggests more diverse keywords used and potentially more informativeness. On the other hand, an entirely irrelevant text may also achieve a high UNI and Length score. Therefore, in addition to UNI, we train Estimated Approval (EA), an external discriminator, whether the generations are valid or not. EA is a binary classification model that is initialized with pre-trained BERT weights (Devlin et al., 2018). We use sub-tasks B and C datasets to fine-tune our EA classifier to estimate the $P(b|X^i, Y'^i)$, where $Y'^i \in \cup(\mathbf{Y}, \hat{\mathbf{Y}})$. The EA score indicates the average number of explanations generated by the language model that are valid explanations according to the EA.

4.2 Analysis

Table 2 summarizes the results obtained by the various models that we have tried for this task. The difference between *UNION w/o CoS-E*, *UNION w/o OpenBook*, *UNION w/o OCMS*, and *UNION* is only the training datasets while the architecture for all the model remains the same. The perplexity, EA score, and the average length of the generations by *UNION w/o CoS-E* are better than *UNION w/o OpenBook*.

It is important to note that the CoS-E dataset is an open-ended common sense question answering. The ComVE dataset, however, has been constructed by annotators who were influenced by ConceptNet to generate false statements. The ConceptNet is a semantic graph of commonsense knowledge developed from the OMCS dataset. Thus, training the UNION model with the OMCS dataset leads to better results than the *w/o OCMS model*. The OpenBook dataset is related to scientific facts, which is similar to the OMCS dataset. Thus, the *UNION w/e CoS-E* performs better than *UNION w/o OpenBook* while the UNION model performs better than all the other models. The initial model we submitted to the SemEval2020 Task 4 is *UNION w/o OCMS*. It achieved a BLEU score of 15.7 and a human evaluation score of 2.10 (Ranked 1st). Later, when training the model with the OMCS dataset, the BLEU score of the model increased by 0.7. Examples of generated explanations from all models are provided in appendix A.

5 Conclusion

In this research, we propose a UNION model for multi-tasking on a couple of commonsense related datasets, which helped our UNION model to achieve state-of-the-art in the ComVE by achieving the

[2] https://github.com/kpu/kenlm

highest human evaluation score. In addition, we propose several auxiliary metrics to better evaluate different models for commonsense response generation tasks without human evaluations. In the future, we would like to explore other possible areas of commonsense reasoning like quantitative reasoning, logic puzzles, and visual common sense reasoning.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Richard E Fikes and Nils J Nilsson. 1971. Strips: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3-4):189–208.

David Gunning. 2018. Machine common sense concept paper. *arXiv preprint arXiv:1810.07528*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

John McCarthy. 1960. *Programs with common sense*. RLE and MIT computation center.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Erik T Mueller. 2014. *Commonsense reasoning: an event calculus based approach*. Morgan Kaufmann.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. *arXiv preprint arXiv:1805.06939*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*.

Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 1223–1237. Springer.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. *arXiv preprint arXiv:1906.00363*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

A Examples of generations

False statement : The Pokemon caught Ash
Baseline : a pokemon is not a bird
Baseline + MTL : pokemon are not animals
UNION w/o Cose : a pokeball is used to catch a pokemons
UNION w/o OpenBook : pokemon are not alive
UNION w/o OCMS : pokemon are not alive
UNION : Ash is a human and cannot be caught by a pokemon.

False statement : You should iron your clothes before you wash them
Baseline : ironing clothes is not the same as washing
Baseline + MTL : ironing clothes is not a good thing to do
UNION w/o Cose : ironing clothes before washing them will make them dirty.
UNION w/o OpenBook : ironing clothes before washing them will make them dirty.
UNION w/o OCMS : ironing your clothes before washing them will make them cleaner
UNION : you should wash your clothes before you iron them

False statement : I was on date with lion yesterday.
Baseline : lion was not on date with girl
Baseline + MTL : lion is not a date.
UNION w/o Cose : lion is not a human to be on date with.
UNION w/o OpenBook : lion is a wild animal and cannot be on a date.
UNION w/o OCMS : lion is not a human to date.
UNION : Lions are wild animals and are not meant to be on dates.

False statement : He broke his leg and went to get a burger
Baseline : burgers are not dangerous
Baseline + MTL : burgers are not broken.
UNION w/o Cose : you don't get a burger from breaking your leg.
UNION w/o OpenBook : a burger is not a medical treatment
UNION w/o OCMS : burgers are not used to heal broken legs.
UNION : Break your leg and you will get a cast.

False statement : People use cars to fly.
Baseline : cars are not the most efficient
Baseline + MTL : cars are not used for flying
UNION w/o Cose : cars don't have the ability to fly.
UNION w/o OpenBook : cars don't fly.
UNION w/o OCMS : cars cannot fly.
UNION : Cars are not designed to fly.