# Commonsense mining as knowledge base completion? A study on the impact of novelty

**Stanisław Jastrzębski**[*]
Jagiellonian University
stanislaw.jastrzebski@uj.edu.pl

**Dzmitry Bahdanau**
MILA
Université de Montréal

**Seyedarian Hosseini**
MILA
Université de Montréal

**Michael Noukhovitch**
MILA
Université de Montréal

**Yoshua Bengio**[†]
MILA
Université de Montréal

**Jackie Chi Kit Cheung**
MILA
McGill University

## Abstract

Commonsense knowledge bases such as ConceptNet represent knowledge in the form of relational triples. Inspired by recent work by (Li et al., 2016), we analyse if knowledge base completion models can be used to mine commonsense knowledge from raw text. We propose *novelty* of predicted triples with respect to the training set as an important factor in interpreting results. We critically analyse the difficulty of mining novel commonsense knowledge, and show that a simple baseline method outperforms the previous state of the art on predicting more novel triples.

## 1 Introduction

Many natural language understanding tasks require commonsense knowledge in order to resolve ambiguities involving implicit assumptions. Collecting such knowledge and representing it in a reusable way is thus an important challenge. There exist several commonsense knowledge bases maintained by experts (CyC) or acquired by crowdsourcing (ConceptNet) which represent commonsense knowledge as relational triples (e.g., *("pen", "UsedFor", "writing")*) (Liu and Singh, 2004). Automatic mining of commonsense knowledge, the focus of this work, aims to improve the coverage of such resources.

One common way of improving the coverage of knowledge bases is through knowledge base completion (KBC), which can be formalized as predicting the existence of edges between (usually) pre-existing nodes in the graph. Recent work by Li et al. (2016) approached commonsense mining as a KBC task. Their method mines candidate triples from Wikipedia and reranks the triples with a KBC model in order to extend ConceptNet.

The goal of this paper is to investigate why recent systems such as the above achieve good performance, and understand their potential for mining commonsense. We approach it by breaking down the previously reported aggregate results into the cases in which models perform well or poorly. We focus in particular on the issue of the *novelty* of model predictions with respect to the triples in the training set. For example, a triple predicted by a system could be correct because it generates output with a slightly different wording or morphological inflection (e.g., *("fish", "AtLocation", "water")* from *("fish", "AtLocation", "in water")*), or it could be correct because it exhibits some degree of semantic generalization (e.g., *("fish", "IsCapableOf", "swimming")* from *("fish", "AtLocation", "in water")*). Arguably, the former could be handled by better standardization of data set formats or more comprehensive model pre-processing, whereas the latter presents an example of genuine commonsense inference and novelty. This analysis is especially important for commonsense mining because of the diversity of the entities, relations, and linguistic expressions thereof in current datasets.

The contribution of this paper is two-fold. First, we test if the KBC task as it is set up in recent work can gauge a model's ability to mine novel commonsense (i.e. find novel commonsense facts based on some resource). We observe the contrary. We present a model that performs poorly on KBC but matches the best model on the task of mining novel commonsense (evaluated by re-ranking extracted candidate triples from Wikipedia). We then examine the cause of this discrepancy, and

[*] Work partially done as intern in MILA
[†] CIFAR Senior Fellow

*Proceedings of the Workshop on Generalization in the Age of Deep Learning*, pages 8–16
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

find that around 60% of triples in the KBC test set used by Li et al. (2016) are minor rewordings of existing triples in the training set. This suggests that controlling for the novelty of triples in both KBC and Wikipedia evaluation is needed.

Second, we present a reassessment of previous methods in which we control the dataset for novelty, extending the results of Li et al. (2016). We introduce a simple automated novelty metric and show that it correlates with human judgment. We then show that the performance of most models on both KBC and Wikipedia triple reranking drops drastically when we evaluate them examples that are genuinely new according to our metric. Finally, we demonstrate that a simple baseline model that does not model all interactions between elements in a triple performs surprisingly well on both KBC and reranking when we focus on novel triples.

## 2 Related work

Knowledge extraction from text corpora is a vast research area (Banko et al., 2007; Mitchell et al., 2015), yet work that targets commonsense knowledge specifically are comparatively rare (Gordon, 2014). Our focus is on the specific approach to mining commonsense knowledge by casting it as a KBC task, as in Li et al. (2016); Forbes and Choi (2017).

Knowledge base completion (KBC) is a method to improve coverage of knowledge base by predicting non-existing edges between nodes (Nickel and Tresp, 2011; Socher et al., 2013). A common modeling approach to KBC is to embed nodes and the edge into a common representation space, followed by a simple prediction model (Socher et al., 2013).

Recently, Dettmers et al. (2017) observed that some KBC benchmarks have test set triples that are simply inversions of triples in their training sets. Our work draws attention to a related issue in commonsense KBC. Additionally, we find that simple baseline models achieve strong performances in our setting, in agreement with other studies of KBC Joulin et al. (2017); Kadlec et al. (2017).

In Angeli and Manning (2013), triple retrieval based on distributional similarity is used to complete ConceptNet. Our procedure for determining the novelty of the triple is similar to methods used in that work, but we apply it only in the context of evaluation.

## 3 Completion vs Mining

Our goal in this section is to analyse the relation between KBC and commonsense mining tasks following setup of Li et al. (2016).

### 3.1 Models

All our models take $(h, r, t)$ triples as inputs, where $h$ and $t$ are sequences of words representing concepts and $r$ is a relation from the ConceptNet schema, and output the probability of the triple to be true. Following Li et al. (2016), we embed $h$ and $t$ by computing the sums $\mathbf{h}$ and $\mathbf{t}$ of the respective word vectors.

Levy et al. (2015) showed that in the context of predicting the hypernymy relation using only head or only tail can be a strong baseline. To better understand how complex reasoning is needed for both KBC and mining tasks, we similarly consider the two following models, which make strong simplifying assumptions about the dependencies between elements in a triple. The **Factorized** model uses only two-way interactions to compute the triple score:

$$\begin{aligned} s(\mathbf{h},\mathbf{r},\mathbf{t}) = \alpha\langle \mathbf{Ah}+\mathbf{b_1}, \mathbf{Bt}+\mathbf{b_2}\rangle \\ +\beta\langle \mathbf{Ar}+\mathbf{b_1}, \mathbf{Bt}+\mathbf{b_2}\rangle \\ +\gamma\langle \mathbf{Ar}+\mathbf{b_1}, \mathbf{Bh}+\mathbf{b_2}\rangle, \end{aligned} \tag{1}$$

where $\mathbf{h}$, $\mathbf{r}$, and $\mathbf{t}$ are $d_1$ dimensional embeddings of head, relation and tail, $\mathbf{A}, \mathbf{B}$ are $d_1 \times d_2$ matrices, $\mathbf{b_1}, \mathbf{b_2}$ are $d_2$ dimensional biases, and $\alpha$, $\beta$, $\gamma$ are learned scalars. The **Prototypical** model is similar, but considers only the head-to-relation and tail-to-relation terms (first and third terms in Eq. 1).

We compare the two new models with the best model from Li et al. (2016), a single hidden layer **DNN**. In that model, the triple score is computed as:

$$\begin{aligned} u(\mathbf{h},\mathbf{t}) = \phi(\mathbf{Ah}+\mathbf{Bt}+\mathbf{b_1}) \\ s(\mathbf{h},\mathbf{r},\mathbf{t}) = \mathbf{W}u(\mathbf{h},\mathbf{t})+\mathbf{b_2}, \end{aligned} \tag{2}$$

where $\phi$ is a nonlinearity, $\mathbf{A}$, $\mathbf{B}$ are $d_1 \times d_2$ matrices, $\mathbf{b_1}$ is a $d_2$ dimensional bias, $W$ is a $d_2$ dimensional vector and $\mathbf{b_2}$ is a scalar. Additionally, we compare against **Bilinear** of Li et al. (2016)[1]. Bilinear model computes the triple score as:

$$s(\mathbf{h}, r, \mathbf{t}) = \mathbf{h}^{\mathbf{T}}\mathbf{M}_r\mathbf{t}, \tag{3}$$

[1] It is the only model evaluated against the Wikipedia ranking task in Li et al. (2016).

where $\mathbf{M}_r$ is a $d_1 \times d_1$ dimensional matrix, separate for each relation in the dataset. All models' scores are fed into a sigmoid function in order to compute the final prediction.

### 3.2 Setup

KBC models are trained using $100,000$ triples from ConceptNet5 (Speer and Havasi, 2012) that were extracted from the Open Mind Common Sense (OMCS) corpus (Speer and Havasi, 2012). For evaluation, we consider two ways to split the dataset: a random split, as well as the confidence-based split proposed by (Li et al., 2016), which uses triples with the highest ConceptNet confidence scores as a test set[2]. Following Li et al. (2016) negative examples are sampled by randomly swapping head, tail or relation component of each triple. The cross-entropy loss is used, and models are evaluated using F1 score[3]. All models are initialized using skip-gram embeddings that were pretrained on the OMCS corpus.

The commonsense mining task is based on a set of $1.7M$ extracted candidate triples from Wikipedia by Li et al. (2016). The extracted triples are ranked using a KBC model, and the top of the ranking is manually evaluated. We will refer to the experiments in which we rerank external candidate triples as *mining* experiments.

We found that similar hyperparameters and optimization methods work well across the models. We use $1,000$ hidden units, and apply L2 regularization with a weight of $10^{-6}$ to the word embeddings. All models are optimized using Adagrad (Duchi et al., 2010) with a learning rate $0.01$ and batch sizes of 200 (DNN) and 600 (Factorized and Prototypical). In Section 3.3, we compare against the scores of a Bilinear model provided by Li et al. (2016). Experiments are performed using Keras (Chollet et al., 2015) and TensorFlow (Abadi et al., 2015).

### 3.3 Comparison of KBC and Wikipedia evaluations

First, we directly test if the performance of a model on the KBC task is predictive of its performance on the mining task. We follow the mining evaluation protocol from (Li et al., 2016): we rank triples by assigned scores and manually evaluate the top 100 resulting triples on a scale from 0 (nonsensical) to 4 (true statement). We re-evaluate their model against our baselines and find that the knowledge base completion task is a poor indicator of performance on Wikipedia. Even though the Factorized and Prototypical models achieve the same or much worse score than DNN on the KBC task (see the first row of Table 1), their mining performance on the top 100 triples is better (than both DNN and Bilinear), see Table 2. Triples were scored by two students and scores were averaged, with $0.81$ Pearson correlation and $0.48$ kappa inter-annotator agreement.

| Novelty \ Model | DNN | Factorized | Prototypical |
|---|---|---|---|
| Entire | $\mathbf{0.892}$ | $0.890$ | $0.794$ |
| $\leq 33\%$ | $\mathbf{0.950}$ | $0.922$ | $0.911$ |
| $(33\%, 66\%]$ | $\mathbf{0.920}$ | $0.898$ | $0.839$ |
| $\geq 66\%$ | $0.720$ | $\mathbf{0.821}$ | $0.574$ |

Table 1: F1 scores on Li et al. (2016) confidence-based test set. F1 score is reported on each bucket (based on the percentile of triple novelty) and the entire test set.

| | Bilinear | Factorized | Prototypical | DNN |
|---|---|---|---|---|
| Wikipedia | 2.04 | **2.61** | 2.55 | 2.5 |

Table 2: Average human assigned score (from 1 to 5) of the top 100 Wikipedia triples ranked by baselines compared to DNN and Bilinear from Li et al. (2016).

### 3.4 Novelty of triples

We hypothesize that the discrepancy reported in Section 3.3 is due to a strong overlap of the training and testing sets in the KBC setup of Li et al. (2016). We perform a human evaluation of the novelty of the triples in the three test sets with respect to the $100,000$ ConceptNet training set used. The first is the confidence-based test set used in Li et al. (2016). We compare it with a random subset of ConceptNet. Finally, we consider a sample of 300 triples from the top $10,000$ triples of Wikipedia dataset ordered by the Bilinear model.

For each triple in the three datasets, we fetch the five closest neighbours using word embedding distance and categorize them into five categories based on the closest triple found in the training set: *"same relation and minor rewording"* (1), *"dif-*

[2] We note that random test set consists of worse quality triples than confidence-based split. However, the latter leads to a serious bias in evaluation. We leave addressing this tradeoff for future work.

[3] The threshold is selected based on a separate development set, as in (Speer and Havasi, 2012).

*ferent relation and minor rewording"* (2), *"same relation and related word"* (3), *"different relation and related word"* (4), *"no directly related triple"* (5). We ignore a small percentage of triples that are not describing commonsense knowledge, as well as false triples (some in the random subset, and a large percentage in the Wikipedia dataset).

To give a better intuition, we provide example triples for the confidence-based split of Li et al. (2016). In Category 1 (defined as "same relation and minor rewording"), we find *("egg", "IsA", "food")*, which has a close analog in the training set: *("egg", "IsA", "type of food")*. An example of a test triple in Category 3 (defined as *"different relation and related word"*) is *("floor", "UsedFor", "walk on")*, which has a corresponding triple in the training set *("floor", "UsedFor", "stand on")*. In the Appendix, we provide more examples of triples from each category.

As shown in Table 3, we observe that approximately 87% examples in the confidence-based test set fall into the first or second category, while these categories constitute only 19% of the considered subset of the Wikipedia triples (even after filtering out false triples). We argue that not controlling for the novelty of triples might introduce hard-to-predict biases in the evaluation.

Finally, to understand the effects of using the confidence-based split, we also re-evaluate models on a random split. We observe that scores are consistently lower than on the confidence-based split (compare the first rows of Tables 1 and 4). Interestingly, the overall performance of the DNN model degrades the most (absolute difference in F1 score 9%), compared to Prototypical (4%) and Factorized (7%).

## 4 Evaluation using novelty metric

Motivated by the described similarity of train and test sets in the KBC task, we shift our attention to re-evaluating models on datasets controlled for novelty, extending results of Li et al. (2016). We consider the same tasks as in Sec. 3: ConceptNet5 completion task and commonsense mining task based on Wikipedia triples.

### 4.1 Automatically measuring novelty

To approximate novelty, we use word embeddings (computed over the OMCS corpus) to calculate distance $d(a, b) = ||head(a) - head(b)||_2 + ||tail(a) - tail(b)||_2$, where head and tail are represented by the average of word embeddings. Such a formulation is related to the concept of *paradigmatic* similarity (Sahlgren, 2006), and word embedding-based distance can approximate paradigmatic similarity (Sun et al., 2015). Two words are paradigmatically similar if one can be replaced for the other, while maintaining syntactical correctness of the sentence (e.g. *"The wolf/tiger is a fierce animal*"). We observe that many trivial test triples are characterized by the existence of a triple in the training set that only differs by such substitutions.

| Novelty<br>Dataset | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Wikipedia | 14% | 5% | 17% | 8% | 44% |
| Confident | 65% | 22% | 4% | 4% | 2% |
| Random | 21% | 10% | 16% | 3% | 29% |

Table 3: Human assigned novelty categories to triples from 3 different test datasets. High quality triples are usually trivial. Each column reports percentage of triples in each category ordered by novelty. Category 1 corresponds to *"same relation and minor rewording"*. Category 5 corresponds to *"no directly related triple"*.

We observe that the proposed distance metric is correlated with human assigned novelty scores (from Sec. 3.4). On the considered datasets Pearson correlation between automatic novelty score and human assigned novelty score is 0.22 to 0.47, with p-values between 0.03 and 0.004. We acknowledge that the automated metric is simplistic, for instance it underperforms for the triples containing rare words or long phrases. Nevertheless, the metric enables detecting a substantial portion of trivial triples (e.g. morphological variations), and we leave for future work developing better measures of novelty.

Using the introduced metric, we can partially explain the inconsistency in the performance of Prototypical and Bilinear models between KBC and mining Wikipedia. We note that the top of the ranking on Wikipedia consists of mostly very far (novel) triples (Figure 1), while KBC confidence-based test set is mostly composed of trivial triples (as argued in Section 3.4).

### 4.2 Novelty-binned evaluation of KBC

We now re-evaluate the KBC models using our proposed novelty metric. First, we examine the performance on different subsets of the

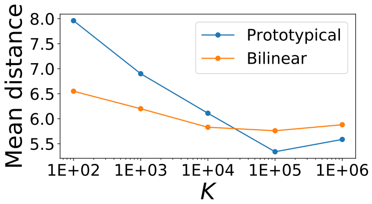

Figure 1: Mean embedding distance (y axis) of top $K$ (x axis) of triples in Wikipedia dataset for Bilinear (orange) and Prototypical (blue).

| Model<br>Novelty | DNN | Factorized | Prototypical |
|---|---|---|---|
| Entire | 0.809 | **0.822** | 0.755 |
| $\leq 33\%$ | **0.883** | 0.874 | 0.866 |
| $(33\%, 66\%]$ | 0.809 | **0.812** | 0.758 |
| $\geq 66\%$ | 0.725 | **0.731** | 0.674 |

Table 4: F1 scores on random split. F1 score is reported on each bucket (based on the percentile of triple novelty) and the entire set.

confidence-based split of ConceptNet5. Specifically, we split the confidence-based test set into 3 buckets, according to 33% (1.93 distance) and 66% (2.80 distance) quantile of distance to the training set. Second, we run a similar experiment but on a random split of the training set (bucket thresholds at 2.1 and 2.95). Results are reported in Tables 1 and 4.

As expected, the performance of models degrades quickly across buckets. The performance on the farthest bucket drops from 10 to 20% F1 score with respect to the performance on the closest bucket. We observe that the Factorized model achieves the strongest performance on the farthest bucket.

### 4.3 Novelty-binned evaluation on Wikipedia

Similar to Section 4.2, we analyse splitting candidate triples for the mining task using our novelty metric. We split the Wikipedia dataset into 3 buckets based on 33% (3.21 distance) and 66% (4.22 distance) quantiles of distance to the training set, and we manually score the top 100 triples in each bucket on the same scale from 1 to 5.

As in Section 4.2, we note a degradation of performance across buckets for all models (from 1.06 to 0.32 mean human assigned score) and again the Factorized model achieves the best performance on the farthest bucket (mean score 2.26 compared to 1.63 and 1.41). The Factorized model outperforms DNN on all buckets despite being a simpler model, which we hypothesize is due to DNN being more prone to overfitting.

| Model<br>Novelty | DNN | Factorized | Prototypical |
|---|---|---|---|
| $\leq 33\%$ | 2.47 | **2.58** | 2.33 |
| $(33\%, 66\%]$ | 2.34 | **2.41** | 2.24 |
| $\geq 66\%$ | 1.41 | **2.26** | 1.63 |

Table 5: Novelty based evaluation of quality of mined triples from Wikipedia dataset. Triples are scored by humans on scale from 1 to 5.

## 5 Conclusions

Mining genuinely novel commonsense is a challenging task, and training successful models will require large training sets (e.g. ConceptNet) and principled evaluation. We critically assessed the potential of KBC models for mining commonsense knowledge, and proposed several first steps towards a more principled evaluation methodology. Future work could focus on developing better novelty metrics, and developing new regularization techniques to better generalize to novel triples.

## Acknowledgments

SJ was supported by Grant No. DI 2014/016644 from Ministry of Science and Higher Education, Poland. Work at MILA was funded by NSERC, CIFAR and Canada Research Chairs.

## References

Martín Abadi et al. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. https://www.tensorflow.org/.

Gabor Angeli and Christopher Manning. 2013. Philosophers are mortal: Inferring the truth of unseen facts. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Sofia, Bulgaria, pages 133–142. http://www.aclweb.org/anthology/W13-3515.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web. In *IJCAI*. volume 7, pages 2670–2676. http://www.aaai.org/Papers/IJCAI/2007/IJCAI07-429.pdf.

François Chollet et al. 2015. Keras. https://github.com/keras-team/keras.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2017. Convolutional 2d knowledge graph embeddings. *CoRR* abs/1707.01476. http://arxiv.org/abs/1707.01476.

John Duchi, Elad Hazan, and Yoram Singer. 2010. Adaptive subgradient methods for online learning and stochastic optimization. Technical Report UCB/EECS-2010-24, EECS Department, University of California, Berkeley. http://www2.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-24.html.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. pages 266–276. https://doi.org/10.18653/v1/P17-1025.

Jonathan Gordon. 2014. *Inferential Commonsense Knowledge from Text*. Ph.D. thesis.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Maximilian Nickel, and Tomas Mikolov. 2017. Fast linear model for knowledge graph embeddings. *CoRR* abs/1710.10881. http://arxiv.org/abs/1710.10881.

Rudolf Kadlec, Ondrej Bajgar, and Jan Kleindienst. 2017. Knowledge base completion: Baselines strike back. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*. pages 69–74. https://aclanthology.info/papers/W17-2609/w17-2609.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 970–976.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1445–1455. https://doi.org/10.18653/v1/P16-1137.

H. Liu and P. Singh. 2004. Conceptnet: A practical commonsense reasoning tool-kit. *BT Technology Journal* 22(4):211–226. https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d.

T. Mitchell, W. Cohen, E. Hruscha, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohammad, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. 2015. Never-ending learning. In *AAAI*. : Never-Ending Learning in AAAI-2015. http://www.cs.cmu.edu/~wcohen/pubs.html.

Maximilian Nickel and Volker Tresp. 2011. A three-way model for collective learning on multi-relational data. In *In Proceedings of the 28th Intl Conf. on Mach. Learn*. Citeseer.

Magnus Sahlgren. 2006. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Ph.D. thesis, Stockholm University, Stockholm, Sweden.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning With Neural Tensor Networks for Knowledge Base Completion. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 926–934.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2015. Learning word representations by jointly modeling syntagmatic and paradigmatic relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 136–145. http://aclweb.org/anthology/P15-1014.

## A Example triples

In this Appendix we report randomly picked examples from human assigned novelty categories considered in the paper for each of the 3 datasets. Due to large size of the training set, instead of showing all triples from train set to human annotator, we show only 5 closest using embedding based distance. A triple is classified as belonging to the given category if *at least one* of the retrieved triples is sufficiently related. For example, if for *("egg", "IsA", "food")* we find triple *("egg", "IsA", "type of food")* in the top 5 closest examples, we categorize it as belonging to the first category ("same rel, rephrase").

### A.1 Confidence-based split

In this Section we report examples for novelty category from Confidence-based split dataset. For each example we include the 5 examples that were shown to the human annotator, ordered by closeness according to our word embedding metric.

#### A.1.1 “same rel, rephrase“

- (egg, IsA, food) : *(egg, UsedFor, food), (egg, HasProperty, good for food), (egg, IsA, type of food), (egg, HasProperty, good for you), (egg, AtLocation, omletts),*
- (book, AtLocation, classroom) : *(lot book, AtLocation, classroom), (physic, AtLocation, classroom), (teacher aide, AtLocation, classroom), (desk and chair, AtLocation, classroom), (test paper, AtLocation, classroom),*
- (dog, CapableOf, be pet) : *(dog, CapableOf, be great pet), (dog, CapableOf, be loyal pet), (dog, CapableOf, be over-fed), (dog, IsA, good pet), (dog, NotDesires, be with cat),*

#### A.1.2 “different rel, rephrase“

- (window, MadeOf, glass) : *(window, HasProperty, make of glass), (window, DefinedAs, glass that be stick to window frame), (abottle, MadeOf, glass), (window, UsedFor, look out of), (window, UsedFor, look inside),*
- (bury cat, HasSubevent, dig hole) : *(bury cat, HasFirstSubevent, dig hole), (bury cat, HasSubevent, dig), (bury cat, HasFirstSubevent, dig grind), (bury cat, UsedFor, when your cat be dead), (bury cat, HasPrerequisite, make sure it be dead),*
- (bridge, UsedFor, cross river) : *(bridge, CapableOf, cross river), (bridge, UsedFor, cross sometihng), (bridge, UsedFor, cross water), (bridge, UsedFor, cross over), (bridge, ReceivesAction, find over river),*

#### A.1.3 “same rel, similar word“

- (cat, CapableOf, hunt mouse) : *(cat, CapableOf, hunt lizard), (cat, NotCapableOf, like mouse), (cat, UsedFor, kill mouse), (cat, CapableOf, kill mouse), (cat, Desires, eat mouse),*
- (pilot, CapableOf, land airplane) : *(pilot, CapableOf, carsh airplane), (pilot, CapableOf, land taildragger), (pilot, CapableOf, work in airplane), (pilot, CapableOf, land), (pilot, AtLocation, airplane),*
- (play sport, HasSubevent, run) : *(play baseball, HasSubevent, run), (play frisbee, Causes, run), (do some exercise, HasSubevent, run), (horse jump high when they, HasProperty, run), (go for run, HasSubevent, run),*

#### A.1.4 “different rel, similar word“

- (statue, AtLocation, museum) : *(statue, ReceivesAction, see in museum), (statue, IsA, example of art), (statue, UsedFor, imortalize someone), (statue, HasProperty, hard to create), (statue, CapableOf, be beautiful),*
- (son, PartOf, family) : *(son, IsA, member of family), (man and his daughter, IsA, family), (son, DefinedAs, child of parent), (son, AtLocation, his home), (son, IsA, male kid of his parent),*
- (internet, UsedFor, research) : *(internet, IsA, amaze research tool), (go on internet, UsedFor, research), (internet, IsA, research project of darpa), (internet, UsedFor, do research or chat), (internet, HasA, lot of information),*

#### A.1.5 “no directly related triple“

- (clerk, CapableOf, stock shelve) : *(clerk, CapableOf, be bag grocery), (clerk, CapableOf, price item), (clerk, CapableOf, bag grocery), (clerk, CapableOf, enter data), (clerk, AtLocation, at hotel),*
- (human, HasA, five finger on each hand) : *(human, HasA, five toe on each foot), (human, HasA, arm hand finger fingernail and lunula), (human, HasA, two hand), (human, CapableOf, write with right hand), (human, CapableOf, stand on two leg),*
- (cat, CapableOf, corner mouse) : *(cat, NotCapableOf, like mouse), (cat, CapableOf, kill mouse), (cat, UsedFor, kill mouse), (cat, UsedFor, keep mouse away), (cat, AtLocation, petstore),*

### A.2 Random split

In this Section we report examples for novelty category from Random split dataset. For each example we include the 5 examples that were shown to the human annotator, ordered by closeness according to our word embedding metric.

### A.2.1 “same rel, rephrase“

- (coffee mug, AtLocation, cupboard) : *(mug, AtLocation, cupboard), (coffee cup, AtLocation, cupboard), (tea cup, AtLocation, cupboard), (cup and plate, AtLocation, cupboard), (can of soup, AtLocation, cupboard),*

- (man, IsA, person) : *(man, IsA, male person), (egoistic person, IsA, person), (woman, IsA, person), (child, InheritsFrom, person), (child, IsA, person),*

- (bookshelf, IsA, for store book) : *(bookshelf, UsedFor, store book), (bookshelf, UsedFor, display and store read material), (bookshelf, UsedFor, hold and organize book), (bookshelf, UsedFor, organize book), (bookshelf, UsedFor, display book),*

### A.2.2 “different rel, rephrase“

- (hear sing, HasSubevent, listen) : *(hear sing, HasFirstSubevent, listen), (hear sing, HasPrerequisite, listen), (hear, HasPrerequisite, listen), (hear music, HasPrerequisite, listen), (hear music, HasSubevent, listen),*

- (procreate, HasPrerequisite, find mate) : *(procreate, HasFirstSubevent, find mate), (procreate, Causes, have to raise your grandchild), (procreate, HasFirstSubevent, form will to do so),*

- (go outside for even, MotivatedByGoal, see star) : *(go outside for even, HasSubevent, that you see star), (go to film, UsedFor, see star), (go outside for even, UsedFor, look at star), (go outside for even, MotivatedByGoal, you have date), (go outside for even, UsedFor, get out of house),*

### A.2.3 “same rel, similar word“

- (aluminum, IsA, metal) : *(aluminum, IsA, material), (safety-pins, MadeOf, metal), (titanium, IsA, metal), (quicksilver, IsA, metal), (plumbum, IsA, metal),*

- (cherry, AtLocation, jar) : *(vegemite, AtLocation, jar), (beet, AtLocation, jar), (toffee, AtLocation, jar), (jellybeans, AtLocation, jar), (moonshine, AtLocation, jar),*

- (u.s president, IsA, political leader) : *(u.s president, IsA, in charge of arm force), (president of something, IsA, it leader), (president, IsA, leader), (president, DefinedAs, leader of american government), (us president, IsA, important political figure),*

### A.2.4 “different rel, similar word“

- (attach case, AtLocation, embassy) : *(attach case, UsedFor, carry paper and book), (attach case, AtLocation, office), (attach case, AtLocation, courtroom), (attache case, AtLocation, businessperson hand), (attache case, CapableOf, hold important document),*

- (catch mumps, Causes, sickness) : *(die, HasSubevent, sickness), (catch mumps, HasSubevent, you have fever), (catch mumps, HasFirstSubevent, get sick), (catch mumps, MotivatedByGoal, be sick), (cold, IsA, sickness),*

- (buy something for love one, Causes, get lay) : *(get in line, MotivatedByGoal, get lay), (have party, UsedFor, get lay), (get pay, UsedFor, get lay), (become inebriate, UsedFor, get lay),*

### A.2.5 “no directly related triple“

- (fall from hot air balloon, CapableOf, kill you) : *(if you drink salt water it, CapableOf, kill you), (drink sea water, CapableOf, kill you), (water, CapableOf, kill you), (lighten, CapableOf, kill you), (pretty thing, CapableOf, kill you),*

- (milk, IsA, part of many food) : *(milk, DefinedAs, product of cow), (milk, ReceivesAction, produce by female cow), (milk, CapableOf, come from cow), (milk, ReceivesAction, make into cheese), (milk, ReceivesAction, create from cow),*

- (some food, ReceivesAction, make from dead animal) : *(some food, HasProperty, good but some be very dissgusting), (some food, IsA, healthy and some be not), (some food, HasProperty, poisonous if prepare improperly), (some food, ReceivesAction, grind before eat), (some food, HasProperty, consider exotic),*

## A.3 Wikipedia

In this Section we report examples for novelty category from Wikipedia dataset. For each example we include the 5 examples that were shown to the human annotator, ordered by closeness according to our word embedding metric.

### A.3.1 “same rel, rephrase“

- (deep snow, IsA, winter) : *(snow, SymbolOf, winter), (snow, AtLocation, winter), (it, IsA, winter), (snowflake, AtLocation, winter), (nice time of year, IsA, winter time),*

- (winter season, HasProperty, cold) : *(winter weather, HasProperty, cold), (in winter it, HasProperty, cold), (snow fall from sky when weather, HasProperty, cold), (stethascopes, HasProperty, cold), (cold weather, Causes, cold),*

- (mathematical logic, HasProperty, logical) : *(mathmatics, HasProperty, logical), (human wish for happiness but happiness, NotHasProperty, logical), (design computer chip, HasPrerequisite, logical think), (write program, HasPrerequisite, logical think), (logic, DefinedAs, set of rule by which axiom can be manipulate to derive true statement),*

### A.3.2 “different rel, rephrase“

- (the house, HasA, room) : *(house, MadeOf, room), (many different way to put furniture, AtLocation, room), (something you find upstairs, IsA, room), (something you find downstairs, IsA, room), (family room, IsA, room),*

### A.3.3 “same rel, similar word“

- (bus system, AtLocation, city) : *(subway system, AtLocation, city), (bus stop, AtLocation, city), (bus, AtLocation, city), (bus shelter, AtLocation, city), (bus station, AtLocation, city),*

- (satellite radio, HasA, channel) : *(tv, HasA, channel), (hear news, HasSubevent, change channel), (watch television, HasSubevent, change channel), (cnn, IsA, television channel), (cnn, IsA, tv channel),*

- (summer, IsA, hotter weather) : *(summer, HasA, more sunshine than winter), (summer, IsA, hot than winter), (summer, IsA, warm than winter), (summer, DefinedAs, season of baseball), (summer, DefinedAs, warm season),*

### A.3.4 “different rel, similar word“

- (liberal democracy, HasProperty, political) : *(democracy, IsA, political system), (liberal democratic party, InstanceOf, japanese political party), (feminism, IsA, political ideology), (libertarianism, IsA, political ideology), (liberalism, IsA, political ideology),*

- (music, UsedFor, musical express) : *(music, CapableOf, be express use musical notation), (music, ReceivesAction, play with musical instrument), (music, ReceivesAction, write with musical symbol), (music, CreatedBy, instrument or human voice), (music, CapableOf, express feel),*

- (the planet, HasA, mass) : *(boston, PartOf, mass), (matter, HasA, mass), (planet plutoi, ReceivesAction, discover by mr), (some planet, HasA, more than one moon), (magnitude of planet, IsA, quantifiable),*

### A.3.5 “no directly related triple“

- (field, HasA, vector potential) : *(field, HasA, plant grow in them), (field, UsedFor, agricultural pursuit), (field, UsedFor, cultivate crop), (field, UsedFor, graze livestock), (field, UsedFor, ride horse),*

- (town, HasA, center of commerce) : *(town, ReceivesAction, compose of many neighborhood), (town, HasProperty, likely to have several cafe), (town, IsA, small than city), (town, DefinedAs, prarie dog community), (town, UsedFor, live in),*

- (divorce, HasProperty, mutual consent) : *(divorce, NotHasProperty, more common than marriage), (divorce, DefinedAs, official end to marriage), (divorce, IsA, fact of life), (divorce, DefinedAs, termination of marriage), (divorce, IsA, when marry couple separate legallyt),*