A Comparison of Two Paraphrase Models for Taxonomy Augmentation

Vassilis Plachouras*
Facebook
1 Rathbone Square, London, UK
vplachouras@fb.com

Fabio Petroni
Thomson Reuters
Corporate Research & Development
30 South Colonnade, London, UK
fabio.petroni@tr.com

Timothy Nugent
Thomson Reuters
Corporate Research & Development
30 South Colonnade, London, UK
tim.nugent@tr.com

Jochen L. Leidner
Thomson Reuters
Corporate Research & Development
30 South Colonnade, London, UK
jochen.leidner@tr.com

Abstract

Taxonomies are often used to look up the concepts they contain in text documents (for instance, to classify a document). The more comprehensive the taxonomy, the higher recall the application has that uses the taxonomy. In this paper, we explore automatic taxonomy augmentation with paraphrases. We compare two state-of-the-art paraphrase models based on Moses, a statistical Machine Translation system, and a sequence-to-sequence neural network, trained on a paraphrase datasets with respect to their abilities to add novel nodes to an existing taxonomy from the risk domain. We conduct component-based and task-based evaluations. Our results show that paraphrasing is a viable method to enrich a taxonomy with more terms, and that Moses consistently outperforms the sequence-to-sequence neural model. To the best of our knowledge, this is the first approach to augment taxonomies with paraphrases.

1 Introduction

Taxonomies are resources for organizing knowledge and are often used in a wide range of tasks such as document classification, search and natural language understanding, among others. Since developing taxonomies is a time consuming process, there has been a significant body of work on their automatic construction. However, even with the application of automatic methods, a taxonomy may not cover all concepts of interest due to issues in bootstrapping the automatic construction, for example the selection of seed terms, the coverage of the data used for mining the taxonomy, or balancing the trade-off between quality and recall.

In this work, we investigate the automatic augmentation of an existing taxonomy using generative paraphrasing. We train a statistical machine translation model and a sequence-to-sequence neural network based model on a subset of the Paraphrase Database (PPDB 2.0). We use the two models to augment an automatically mined taxonomy of risk terms based on (Leidner and Schilder, 2010).

The research questions we address in this work are the following:

- **RQ1** Can the models generate high quality paraphrases for automatically augmenting a taxonomy?
- **RQ2** How much does the coverage of the taxonomy increase?
- **RQ3** Which model is best for generating paraphrases?

We answer these research questions by assessing the quality of the generated risk phrases and quantifying the number of additional sentences that the generated paraphrases match in a large corpus of news articles.

2 Related Work

Paraphrase Generation. Identifying and generating paraphrases has received significant attention, being useful in applications ranging from natural language understanding, to query expansion for example (Madnani and Dorr, 2010; Androutsopoulos and Malakasiotis, 2010).

A number of works treat paraphrase generation as a special case of machine translation, learning to generate paraphrases based on a large number of

*work conducted whilst the author was at Thomson Reuters.

Proceedings of NAACL-HLT 2018, pages 315–320
New Orleans, Louisiana, June 1 - 6, 2018.

aligned sentence pairs from news articles (Quirk et al., 2004), extracting paraphrases from a bilingual parallel corpus (Bannard and Callison-Burch, 2005), or training statistical machine translation models on news headlines (Wubben et al., 2010).

Building on the recent advances in neural networks for machine translation, seq2seq models with attention representing input as a sequence of characters (Hasan et al., 2016), or with more layers and residual connections (Prakash et al., 2016) have been trained to generate paraphrases. Mallinson, Sennrich and Lapata (2017) applied the bilingual pivoting approach (Bannard and Callison-Burch, 2005) with neural machine translation, where the input sequence is mapped to a number of translations in different languages, and then these translations are mapped back to the original language.

Taxonomy Construction & Expansion Since manually creating knowledge structures, such as taxonomies, is a time consuming process, there exist several methods to automate it (Medelyan et al., 2013). Meng et al. (2015) employ techniques for automatically mining taxonomies in combination with crowd-sourcing to achieve greater coverage. Subramaniam, Nanavati and Mukherjea (2010) study the problem of merging one ontology into another one, thus asymmetrically extending one of the taxonomies. Harpy (Grycner and Weikum, 2014) addresses the sparsity of subsumption hierarchy of Patty, a large repository of relational paraphrases (Nakashole et al., 2013). Wang et al. (2014) automatically extend a taxonomy by identifying missing categories and predict the optimal structure based on a hierarchical Dirichlet model. The automatic placement of new concepts in a taxonomy has also been investigated as a shared task in SemEval 2016 (Jurgens and Pilehvar, 2016). However, to the best of our knowledge, there is no work that applies generative paraphrasing to expand a taxonomy.

3 Paraphrase Generation

In this work we approach the task of generating phrasal paraphrases as monolingual translation and we train two state-of-the-art models (Section 3.1) on an existing corpus of English phrasal paraphrases (Section 3.2).

3.1 Models

The two models we train for paraphrase generation are based on Moses (Koehn et al., 2007) and attention-based sequence-to-sequence (seq2seq) neural networks (Bahdanau et al., 2015).

Moses is an open-source implementation of statistical machine translation models. While it supports the use of additional structure such as dependency trees, we focus on phrase-based translation in this work and a tri-gram language model learned from the set of target paraphrases.

The attention-based seq2seq model consists of a bi-directional LSTM encoder and an LSTM decoder which uses an attention mechanism to learn which input words are the most important for each output word.

3.2 Training and Evaluation

For training the paraphrase generation models, we use a subset of the Paraphrase Database (PPDB 2.0) corpus. The PPDB 2.0 data set is a large-scale phrasal paraphrase data set that has been mined automatically based on (Bannard and Callison-Burch, 2005), and refined with machine learning based ranking of paraphrases based on human generated ground-truth and assignment of textual entailment labels for each paraphrase pair. In this work, we used the large pack of lexical (single word to single word) and phrasal (multi-word to single or multi-word) paraphrases[1]. Because the data set was automatically generated, some of the paraphrase pairs are not true paraphrases. In our experiments, we kept only pairs that do not contain numeric digits. We also use the textual entailment labels with which paraphrase pairs in the data set are annotated and keep the pairs labeled as equivalent. We split the remaining data in 757,300 training data points and 39,325 test data points. The splitting is performed by first creating a graph where phrases are nodes and edges exist between the two phrases in a paraphrase pair. In this graph, we identify connected components and we assign all data points within each connected component to either the training or the test sets. This process guarantees independence between the training and the test sets.

To train Moses, we precomputed a tri-gram language from the target phrases in the training data

[1]PPDB 2.0 is made available in packs of increasing size, where each pack contains a list of paraphrases, ordered in decreasing order of the score described in (Pavlick et al., 2015).

set and used the MERT optimizer. To train the seq2seq model, we used a batch size of 256 training samples, 100-unit LSTM cells for both the encoder and the decoder, dropout with keep probability 0.8 at the output of cells, a bidirectional encoder, greedy 1-best search output generation criteria, and an additive attention function (Bahdanau et al., 2015). For representing words, we used 100 dimensional pre-trained GloVe embeddings (Pennington et al., 2014). We trained using the Adam optimizer and a learning rate of 0.001 and let the models train for 200,000 steps (a step is an iteration over a batch of training points).

For evaluation we used the BLEU score (Papineni et al., 2002; Chen and Cherry, 2014). BLEU is calculated on tokenized data using the implementation provided in the *nltk* framework[2] with NIST geometric sequence smoothing. Moses achieved a BLEU score of 0.4098 compared to 0.3156 obtained by the seq2seq model. The difference in BLEU score shows that Moses is substantially better than the seq2seq model for the subset of PPDB 2.0 we used.

4 Taxonomy Augmentation Evaluation

After training the paraphrase generation models, we focus on augmenting the taxonomy of risks. The risk taxonomy has been automatically mined based on the method described in (Leidner and Schilder, 2010) and subsequently has been manually filtered to keep high quality risk terms, resulting in 2,824 terms.

For each term in the risk taxonomy, we apply the two paraphrase generation models to obtain a maximum of top 10 paraphrased risk terms. Figure 1 shows the number of generated paraphrases that are also in the original list of risk phrases. While our end goal is to generate phrases that are not in the original list of phrases, a large number of generations already appearing in the list of high-quality and manually filtered list of risk phrases is an indication of the quality of the paraphrases. As we can see from Figure 1, Moses outperforms with a wide margin the seq2seq model in generating paraphrases already in the taxonomy. Table 1 shows examples of generated paraphrases by Moses and seq2seq.

Furthermore, we have manually annotated the top-1 generated paraphrases that were not already in the original risk taxonomy. Each paraphrased

[2] http://www.nltk.org/

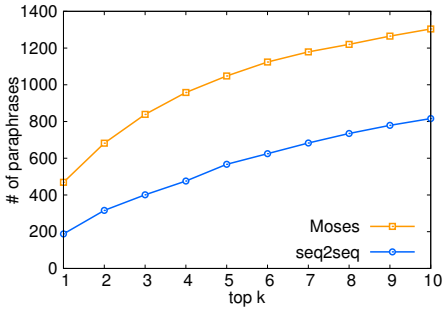

Figure 1: Number of top-k generated paraphrases already in the list of risk phrases.

risk term was annotated as *valid* when it can directly replace the original risk term, *noisy* when the meaning of the paraphrase is close to the meaning of the original term or the paraphrase has additional terms, and *invalid* when the paraphrase is not suitable for substituting the original risk term. Table 2 shows for both models the number of paraphrases that were not in the original taxonomy and that were annotated with a given label.

Even though both the BLEU score and the number of paraphrases that were already in the original risk taxonomy demonstrate that Moses performs better than seq2seq in our setting, we have also looked how often a paraphrase generated with seq2seq was annotated as being better than the paraphrase generated by Moses. For example, this is the case when one model generates a paraphrase that is annotated as *valid* and the other model generates for the input risk term a paraphrase that is annotated as *noisy* or *invalid*. We have observed that in 1211 cases, the paraphrase generated by Moses was better than the paraphrase generated by seq2seq. On the other hand, seq2seq was better only in 58 cases.

We have also looked at the lexical diversity of the generated paraphrases. We define lexical diversity as the fraction of tokens in a paraphrase that were not in the original risk phrase. Table 3 shows that the seq2seq model results in higher lexical diversity than Moses for both the valid and noisy paraphrases.

Finally, we have looked at the number of sentences matched by the original risk phrases and the generated paraphrases in large corpus of approximately 14 million news articles. The original list of risk phrases matches 23,110,506 sentences. As Table 4 demonstrates, the *valid* paraphrases generated by Moses match an additional 5.2M sentences that were not matched by any entry in the

Risk Term	Moses	seq2seq
wind-blown debris	wind-blown rubble	buildings
unexpected entry of competitors	unpredicted entrance of competitors	accident
trafficked people	trafficking in persons	victims of human trafficking
demolished	razed	demolition
committed fraud	fraud committed	the fight against fraud
genetically modified food	gm food	genetically engineered

Table 1: Examples of paraphrases generated by Moses and seq2seq.

Model	Valid	Noisy	Invalid
Moses	1,337	337	327
seq2seq	419	175	2,042

Table 2: Number of generated paraphrases annotated as valid, noisy or invalid.

Model	Diversity (valid)	Diversity (noisy)
Moses	0.5455 (1,337)	0.3952 (337)
seq2seq	0.6991 (419)	0.6969 (175)

Table 3: Lexical diversity of generated valid and noisy paraphrases in terms of fraction of tokens that are not in the original risk phrase.

Model	Valid	Noisy
Moses	5,197,781	1,868,734
seq2seq	1,751,745	749,886

Table 4: The number of sentences from the news archive matching at least one of the generated valid or noisy risk paraphrases, which were not already matched by a risk phrase in the original taxonomy.

original taxonomy, corresponding to an expansion of coverage by 22%. The *valid* paraphrases generated by seq2seq match 1.8M additional sentences, expanding coverage by 7.6%. A smaller increase in coverage can be achieved if we consider *noisy* paraphrases as well, 8.1% for Moses and 3.2% for seq2seq. However, these additional sentences may contain significantly more noise.

Overall, we have seen that the application of paraphrase generation can expand an existing taxonomy of risk terms with high quality phrases, where 67% of the added terms by Moses have been assessed as valid paraphrases (**RQ1**). This has led to an increase of the coverage of the taxonomy by 22% (**RQ2**). The experimental results also demonstrate that Moses outperforms the neural network-based model in this setting (**RQ3**).

5 Discussion

Domain-specific paraphrases. During the annotation of the generated paraphrases by the two models, we have observed a number of cases, which were annotated as *invalid* because the generated paraphrase, although it was grammatically correct and meaningful, it did not correspond to the original term in the domain of risk management. For example, the phrase "screening risk", which refers to risks in the process of performing background checks, was paraphrased to "projection risk" by one of the models. Even though the latter is a grammatically correct phrase, it does not have the same meaning in the context of risk management. Similarly, the word *concentrations* has been replaced by the word *levels* in the phrase "sector concentrations", which may be more appropriate as a replacement in the domain of chemistry. A more appropriate word to replace *level* would be *focus*. To address this issue of domain specific paraphrasing, one possible solution is to use a domain-specific corpus to train the language model used in Moses, or to pre-train the weights of the LSTM cells in the encoder and decoder of seq2seq in the context of a language modelling task (Dai and Le, 2015).

Grammatical diversity. We have quantified lexical diversity as the fraction of new words in the generated paraphrases. Another aspect of diversity, however, is grammatical diversity. For example, it would be interesting to quantify diversity in terms of the number of the classes of paraphrasing phenomena defined by Bhagat and Hovy (2013).

6 Conclusions

In this work we have looked at the problem of automatically augmenting a taxonomy by generating paraphrases of the terms in the taxonomy. Using a subset of PPDB 2.0, a data set of paraphrases, we have trained a statistical machine translation model based on Moses and a second one based on sequence-to-sequence neural network-based mod-

els. Our evaluation results show that Moses outperforms seq2seq in our setting and it augments the taxonomy with 67% of high quality terms, leading to an increase of coverage by 22%.

For future work, we want to investigate the impact of using pre-trained weights to initialize the LSTM cells in the seq2seq model from a language modelling task, as well the grammatical diversity of generated paraphrases.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *J. Artif. Int. Res.* 38(1):135–187.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. ACL '05, pages 597–604.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics* 39:463–472.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. *ACL 2014* page 362.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc., pages 3079–3087. http://papers.nips.cc/paper/5949-semi-supervised-sequence-learning.pdf.

Adam Grycner and Gerhard Weikum. 2014. Harpy: Hypernyms and alignment of relational paraphrases. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, pages 2195–2204. http://www.aclweb.org/anthology/C14-1207.

Sadid A. Hasan, Bo Liu, Joey Liu, Ashequl Qadir, Kathy Lee, Vivek Datla, Aaditya Prakash, and Oladimeji Farri. 2016. Neural clinical paraphrase generation with attention. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*. pages 42–53.

David Jurgens and Mohammad Taher Pilehvar. 2016. Semeval-2016 task 14: Semantic taxonomy enrichment. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, pages 1092–1102. https://doi.org/10.18653/v1/S16-1169.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics, pages 177–180. http://www.aclweb.org/anthology/P07-2045.

Jochen L. Leidner and Frank Schilder. 2010. Hunting for the black swan: Risk mining from text. In *Proceedings of the ACL 2010 System Demonstrations*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACLDemos '10, pages 54–59. http://dl.acm.org/citation.cfm?id=1858933.1858943.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Comput. Linguist.* 36(3):341–387.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. EACL '17, pages 881–893.

Olena Medelyan, Ian H. Witten, Anna Divoli, and Jeen Broekstra. 2013. Automatic construction of lexicons, taxonomies, ontologies, and other knowledge structures. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 3(4):257–279. https://doi.org/10.1002/widm.1097.

R. Meng, Y. Tong, L. Chen, and C. C. Cao. 2015. Crowdtc: Crowdsourced taxonomy construction. In *2015 IEEE International Conference on Data Mining*. pages 913–918. https://doi.org/10.1109/ICDM.2015.77.

Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2013. Discovering semantic relations from the web and organizing them with patty. *SIGMOD Rec.* 42(2):29–34. https://doi.org/10.1145/2503792.2503799.

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*. ACL '02, pages 311–318.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. 2015. Adding semantics to data-driven

paraphrasing. 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 1512–1522. https://doi.org/10.3115/v1/P15-1146.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. EMNLP '14, pages 1532–1543.

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farry. 2016. Neural paraphrase generation with stacked residual lstm networks. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*. COLING '16, pages 2923–2934.

Chris Quirk, Chris Brockett, and William B. Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. EMNLP '04, pages 142–149.

L. V. Subramaniam, A. A. Nanavati, and S. Mukherjea. 2010. Enriching one taxonomy using another. *IEEE Transactions on Knowledge and Data Engineering* 22(10):1415–1427. https://doi.org/10.1109/TKDE.2009.189.

Jingjing Wang, Changsung Kang, Yi Chang, and Jiawei Han. 2014. A hierarchical dirichlet model for taxonomy expansion for search engines. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, New York, NY, USA, WWW '14, pages 961–970. https://doi.org/10.1145/2566486.2568037.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2010. Paraphrase generation as monolingual translation: Data and evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*. INLG '10, pages 203–207.