Predicting Informativeness Of Semantic Triples

Judita Preiss
University of Salford
Salford M5 4WT, United Kingdom
j.preiss@salford.ac.uk

Abstract

Many automatic semantic relation extraction tools extract subject-predicate-object triples from unstructured text. However, a large quantity of these triples merely represent background knowledge. We explore using full texts of biomedical publications to create a training corpus of informative and important semantic triples based on the notion that the main contributions of an article are summarized in its abstract. This corpus is used to train a deep learning classifier to identify important triples, and we suggest that an importance ranking for semantic triples could also be generated.

1 Introduction

Subject-predicate-object triples are used in numerous natural language processing areas, including question answering (e.g. Hristovski et al. (2015)), ontology building (e.g. Du and Li (2020)) and literature based discovery (e.g. Hristovski et al. (2006)). While they can be thought of as representing the minimum unit of semantic expression, there is a large degree of variability in the amount of new (not commonly known) content they convey. On the one hand, they sometimes represent what can be termed background knowledge, for example *"New Zealand - ISA - country"* or *"pharmaceutical services - TREATS - health personnel"*, while on the other, they may describe very specific findings such as *pimobendan TREATS hypertrophic cardiomyopathy* or *LCN2 protein, human - ASSOCIATED_WITH - chronic kidney disease*. We use biomedical publications to test the hypothesis that training data consisting of such, important, triples can be created from abstracts, and train a deep learning algorithm to identify these high importance triples from a list of all triples appearing in a paper. The system could also be adjusted to output a weight instead of a binary decision, allowing for an importance ranking of semantic triples within an article.

The paper begins with an overview of related work in Section 2, the experimental set-up follows in Section 3, with the results and discussion in Section 4 and conclusions drawn in Section 5.

2 Background

A number of tools for automatically extracting semantic relations – (subject, relation, object) triples – from unstructured text exist (Yuan and Yu, 2018). However, as Papadopoulos et al. (2020) point out, the majority of works incorporating these do not perform much pre- or post- processing and therefore include many potentially uninformative triples, and works proposing to extend currently existing collections of semantic relations often speak of extending the set of relations, not refining the relations present (e.g. Koroleva et al. (2020)).

Evaluations of semantic relation extraction systems are often very comprehensive, e.g. Kilicoglu et al. (2020) present a detailed independent evaluation of SemRep – a biomedical domain tuned triple extraction tool – and discover common sources of error for this tool, but such evaluations do not quantify the quality of the triple that is retrieved by the system. It is unclear whether the incorrectly extracted triples are uninformative, or the opposite.

While not phrased as focusing on informative / important triples, existing works often restrict to particular types of relations: Yuan and Yu (2018) evaluate the extraction of health claims, defined as a relation between something that is being manipulated and something that is being measured (e.g. the relation between a substance and a disease). Yadav et al. (2020) restrict to drug-drug interaction, protein-protein interaction, and medical concept relation extraction, while Hope et al. (2021) focus on mechanisms, i.e. activities, functions and

Proceedings of Recent Advances in Natural Language Processing, pages 1124–1129
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_126

causal relations. Such restrictions are likely to increase the overall quality of the remaining triples: removing the *ISA* relation alone eliminates a large quantity of background knowledge. The closest to our work is due to Zhang et al. (2021) who filter out uninformative triples computationally, based on the difference between triples' expected and observed frequencies.

3 Experiment Design

Below, we discuss the two steps needed to explore the hypothesis that a dataset based on abstracts can be used to detect important triples using machine learning: 1) creation of a training corpus, and 2) selection of a deep learning architecture.

3.1 Training Corpus Creation

The CORD-19 dataset (Wang et al., 2020) was chosen for this work due to: 1) scale, the 2021-05-03 version contains 198,347 full articles, 2) availability of extracted text, the dataset contains the text extracted from available full article PDFs, 3) domain, the restricted nature of the dataset allows the application of existing biomedical tools.

3.1.1 Semantic Relation Extraction

Subject-relation-object triples are extracted from all article texts present in the dataset using SemRep (Rindflesch and Fiszman, 2003). Designed for the biomedical domain, the tool extracts triples such as "*imatinib TREATS Gastrointestinal Stromal Tumors*" but with concepts mapped to Unified Medical Language System metathesaurus (UMLS) (Bodenreider, 2004) concept unique identifiers, CUIs (i.e. yielding *C0935989 - TREATS - C0238198* for the example). This addresses the problem of multi-word identification (recognizing *gastrointestinal stomal tumours* rather than merely *tumours*) and word sense disambiguation (distinguishing between occurrences of concepts with multiple meanings, such as COLD, which could - among other options - represent the *common cold* or *chronic obstructive airway disease*).

3.1.2 Identifying Important Triples

To train a machine learning classifier, a training set of important triples is needed. Since an abstract usually summarizes the main findings of an article, we hypothesize that important triples can be considered to be those that appear in both the body and an abstract. It is important to note that the training set of important triples does not need to be complete, i.e. not every important triple from the body needs to be identified. The dataset should be as noise free as possible, and therefore background knowledge triples (which may appear in both the abstract and the body of an article) should not be included. To reduce noise, the following filtering is performed:

- Previously published triples. The construction of positive examples in the training set hinges on the identification of important triple(s). If these triples are defined as those which describe the novel contribution(s) of an article, an identical triple (i.e. contribution) should not have appeared in abstracts prior to the current paper. Therefore triples appearing in SemRep processed Medline (V40, released October 2019, i.e. before the CORD-19 dataset), a vast collection of biomedical abstracts (Lozano-Kühne, 2013), are removed from the dataset.

- Frequent concepts. Some frequent concepts often appear in non important triples, such as:

  - ***therapeutic procedure*** TREATS *disease*
  - *malaise* PROCESS_OF ***patients***
  - *lung* PART_OF ***homo sapiens***

  Since the training set does not need include an annotation for every triple encountered and there is high probability of mis-annotation with triples involving these concepts, triples involving the top 1% of concepts appearing in V40 of SemRep processed Medline are removed. The top 1% includes *patients*, *therapeutic procedure*, *homo sapiens* and other very general terms. Note that this does not mean that the system will be unable to classify triples including these concepts.

In some cases, an identical triple is used both in the abstract and the body of an article, however, when repeated, novel contributions of a paper are sometimes rephrased using (near) synonyms. Therefore a measure of triple similarity needs to be defined. Since the triples are of the format $subject_{CUI}$-$predicate_{word}$-$object_{CUI}$, this measure can be defined on each component (subject, predicate, object) separately. Word (CUI) embeddings represent each word (CUI) as a vector which captures information about the contexts it appears in, therefore yielding similar – close – vectors for synonyms. A triple similarity measure can therefore be implemented based on cui2vec (Beam et al.,

2019) (for subject and object similarity) and GloVe (Pennington et al., 2014) embeddings (for predicate similarity).[1] Similarity between two triples, $cui_{11}-rel_1-cui_{12}$ and $cui_{21}-rel_2-cui_{22}$, is then given by the formula $cs(c2v(cui_{11}), c2v(cui_{21})) + cs(g(rel_1), g(rel_2)) + cs(c2v(cui_{12}), c2v(cui_{22}))$ where cs represents the cosine similarity, $c2v(x)$ the cui2vec vector of x and $g(x)$ x's GloVe vector. As the maximum value for cosine similarity is 1, the triple similarity is a decimal between 0 and 3 inclusive, 0 corresponding to complete lack of similarity between triples and 3 an exact match. For each body-triple, a similarity can be computed between it and each abstract-triple in the same article, with the highest becoming the body-triple's *similarity value*. A threshold can be set on the similarity value to decide which triples are deemed important.

3.2 Deep Learning Algorithm

The machine learning component consists of three parts: 1) feature extraction, 2) architecture selection, and 3) experiment settings.

3.2.1 Feature Extraction

The ability to extract important triples (described in Section 3.1.2) makes it possible to use supervised machine learning approaches to train a classifier. To this end a number of features are extracted for each body-triple.

Frequency based features: 1) the number of times the triple appeared in the body of the article, and 2) the total number of relations within the body of the publication.

UMLS based features: 1) the frequency count of the CUIs in the body triple as extracted from SemRep processed Medline – while the top 1% of CUIs have been discarded, it is believed that CUIs with lower frequencies are more likely to be part of novel contributions, 2) the UMLS source vocabulary of the CUIs – the metathesaurus consists of many different types of biomedical vocabularies and the information pertaining to which one(s) a CUI belongs to can serve to give an overall idea of its category, and 3) the depth of the body triple CUIs within UMLS. For some source vocabularies, a hierarchy is present, allowing the computation of the concept's distance to the root – assuming a concept further away from the root is more likely to be more fine-grained, this feature also investigates whether important triples are more likely to contain more specific CUIs (the shortest path to the root is taken if a concept appears in multiple hierarchies).

Semantics based features: 1) the relation used, 2) the title of the section the body triple appeared in – since the majority of articles in this collection have relatively rigid structure, this was restricted to the commonly prescribed sections such as *introduction*, *background*, *methods* etc, and is based on the hypothesis that a novel contribution of a work is likely to appear in the *discussion* and / or *conclusion* sections, and 3) the rank of the sentence the triple appeared in as ranked by TextRank (Mihalcea and Tarau, 2004). TextRank is a graph based algorithm, often used in summarization, which can be used to order the sentences in an article according to importance, and therefore we hypothesize that a sentence with a low TextRank (high importance) is more likely to yield an important triple.

After performing one hot encoding of the relation feature, this gives 129 features for the 55,745 triples in the dataset.

3.2.2 Architecture Selection

While the similarity value of a body-triple calculated as described in Section 3.1.2 can be predicted directly, initial experiments with regression showed that this is hard to do exactly. The problem was therefore framed as binary classification. In this case, a threshold is set on the similarity value and triples with a value above the threshold are used as positive, important, instances.

Deep learning model is chosen due to its ability to cope with feature dependencies. The model, implemented using Keras, was designed with fully connected (dense) layers of halving sizes with the final layer of size 1. ReLU was used for all layers except the last, where the sigmoid activation function was employed. The loss function was binary entropy and accuracy was used as the metric when classes weren't extremely imbalanced, F_1 was used otherwise. A number of parameters were tuned: 1) the depth of the model (with halving sizes, thus depth one model has a single dense layer of size int(129/2), depth two model has two dense layers of sizes int(129/2), int(129/4), and so on), 2) the number of epochs, 3) dropout, and 4) whether class weights were used.

3.2.3 Experiments

As suggested above – by exploring the use of class weights within the model – the dataset is highly imbalanced with, as expected, the majority of triples

[1] GloVe embeddings were chosen since the predicate words are being compared in isolation.

Similarity value	Buffer band	Majority proportion	Best model Depth	Dropout	Accuracy / F-measure
≥ 3	ON	50	2	0.0	a=72.7
≥ 3	OFF	50	2	0.0	a=67.2
≥ 3	ON	83.3	2	0.0	a=85.2
≥ 3	OFF	83.3	2	0.0	a=84.2
≥ 2	OFF	88.3	3	0.0	f=0.975

Table 1: Performance of informative triple classifier

not appearing in the abstract. The following methods for addressing this bias were explored:

- Using class weights within the deep learning algorithm: this allows more emphasis to be given to the minority class.
- Under-sampling: randomly sampling the majority class such that the number of examples used in training corresponds to a pre-decided ratio. The minority and majority class can be made equal (1:1) but other ratios were explored, making the majority class more frequent but not overpowering.

While all the minority, important, class triples are included in the training set, this does not have to be the case for the majority class. As mentioned above, the triples to include in the minority class are selected by a threshold. However, this can lead to a triple with, say, similarity of 2.5 being included in important triples, while a triple with similarity of 2.499 appearing in the non important triples class. Such small difference may be detrimental to the performance of the machine learning algorithm and a buffer band of similarities between the two classes was also explored. I.e. two thresholds, t_1 and t_2 are set such that $t_1 - t_2 > 0$ and all triples with similarity $>= t_1$ are assigned to the important class while triples with similarity $<= t_2$ are deemed not important.

4 Results And Discussion

A 5-fold cross validation was performed, and each explored model was trained on (a possibly balance adjusted version of) the training portion giving rise to an accuracy or F-measure on the test portion. This allows an average to be computed and the best model to be determined. The results are presented in Table 1: the similarity value refers to the threshold from Section 3.1.2 used to determine which triples are considered important, the buffer band – when on – removes the cases close to the similarity value threshold from training as described in Section 3.2.3, and the majority column represents the percentage of the training dataset attributed to the majority class. The final columns present the hyperparameters of the best model for the specific combination and the average accuracy / F-measure.

With under-sampling, the accuracies for similarities $>= 2$ were all within 2% of the best performance, supporting the hypothesis regarding frequent use of synonyms. To avoid a uniform assignment of the majority class, the F-measure metric (which rewards both precision and recall) is used in models without under-sampling. An F-measure of 1 represents perfect precision and recall, and the highest F-measure achieved is 0.975.

SHapley Additive exPlanations (SHAP) (Lundberg et al., 2018) uses ideas from game theory to explain feature contributions to machine learning decisions. Figure 1 depicts the feature contributions on a randomly selected sample of 100 triples for the best model without under-sampling. Each dot represents a single triple, with the intensity (blue → pink) indicating whether the feature value was low or high. The horizontal position indicates whether the contribution caused the prediction to go up – towards being classified as an important triple – or down. The top three rows show expected results: that high values in the number of relations in the document, very frequently occurring CUIs or relations arising from sentences low in importance ranked by TextRank (giving a high rank) impact the prediction very negatively. Unsurprising positive contributors are: 1) the frequency of the triple in the document: a new contribution may be reiterated in the document, 2) the triple appearing in the conclusion: this often contains a summary of contributions, 3) the triple including the *TREATS* relation: the filtering ensures this is a new triple and being treatment specific, is likely the focus of the work, 4) the triple appearing in the intro-

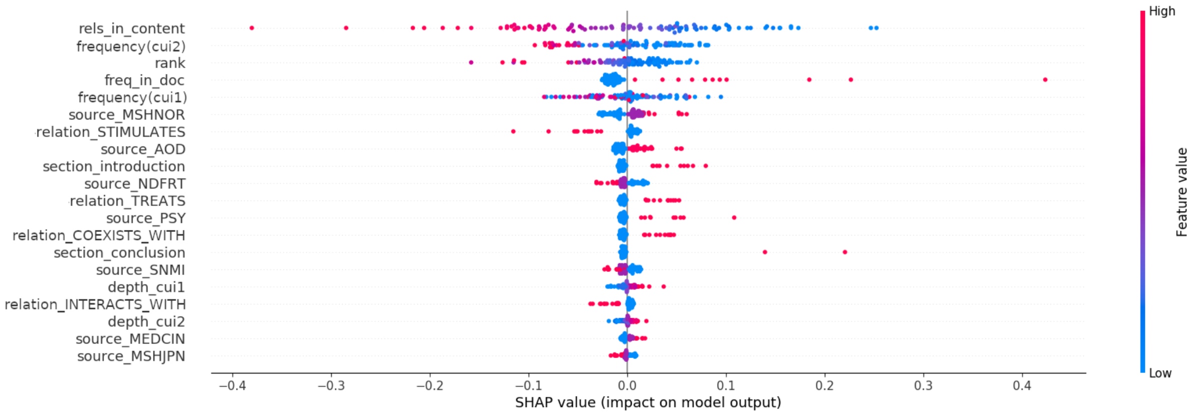

Figure 1: Feature contributions

duction, where the novelties of the work are often highlighted. The contributions of a higher depth value is also, as expected, positive.

Contributions are also due to the CUIs' UMLS source vocabulary (indicated by *source_*). In some cases, these categorize the CUI: for example, AOD (alcohol and other drug thesaurus) and PSY (psychological index terms) are not unexpected. Surprising may be the pair MSHNOR and MSHJPN, representing the Norwegian and Japanese translations of Medical Subject Headings, as they appear to have opposite effect. However, MSHJPN's contribution is very limited, suggesting that its completeness may not match that of MSHNOR.

5 Conclusions And Future Work

We have demonstrated that a dataset of semantic triples created from full articles based on similarity between triples in the body of the text and triples in the abstract can be used to train a deep learning classifier to make predictions about a semantic triple's importance. An analysis of feature contributions was also performed.

While a direct prediction of the similarity score appeared difficult with the quantity of data available, converting the similarity scores into categorical values may be trainable and would provide the basis of a ranking. Again with greater quantity of data, features based on medical subject headings of each CUI could be beneficial indicated by the success of the UMLS source vocabulary features.

The work undertaken was in the biomedical domain based on a tool tuned for biomedical domain grammatical relation extraction. Porting the approach to another domain, where subject-verb-object triples would need to be extracted using a generic grammatical relation extraction algorithm and some features would require re-engineering, would also form an interesting extension of the work.

Acknowledgements

This work was partly supported by the ERDF Greater Manchester AI Foundry grant, and I would like to thank Sunil Vadera in particular for his support. Thanks also go to the anonymous reviewers for ensuring clarity of this paper as well as pointing out some very nice avenues for further work.

References

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan P. Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. 2019. Clinical concept embeddings learned from massive sources of multimodal medical data. *arXiv preprint arXiv:1804.01486*.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

Jian Du and Xiaoying Li. 2020. A knowledge graph of combined drug therapies using semantic predications from biomedical literature: Algorithm development. *JMIR Med Inform*, 8(4):e18323.

Tom Hope, Aida Amini, David Wadden, Madeleine van Zuylen, Sravanthi Parasa, Eric Horvitz, Daniel S. Weld, Roy Schwartz, and Hannaneh Hajishirzi. 2021. Extracting a knowledge base of mechanisms from COVID-19 papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4489–4503.

Dimitar Hristovski, Dejan Dinevski, Andrej Kastrin, and Thomas Rindflesch. 2015. Biomedical question answering using semantic relations. *BMC Bioinformatics*, 16(6).

Dimitar Hristovski, Carol Friedman, Thomas C. Rindflesch, and Borut Peterlin. 2006. Exploiting semantic relations for literature-based discovery. In *Proceedings of the 2006 AMIA Annual Symposium*, pages 349–353.

Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Dongwook Shin. 2020. Broad-coverage biomedical relation extraction with semrep. *BMC bioinformatics*, 21(1):188.

Anna Koroleva, Maria Anisimova, and Manuel Gil. 2020. Towards creating a new triple store for literature-based discovery. In *Trends and Applications in Knowledge Discovery and Data Mining*, pages 41–50.

Jingky Lozano-Kühne. 2013. *MEDLINE*, pages 1218–1218. Springer New York.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411.

Dimitris Papadopoulos, Nikolaos Papadakis, and Antonis Litke. 2020. A methodology for open information extraction and representation from large scientific corpora: The CORD-19 data exploration use case. *Applied Sciences*, 10(16).

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*, pages 1532–1543.

Thomas C. Rindflesch and Marcelo Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathrun Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 open research dataset. ArXiv [Preprint]. 2020 Apr 22:arXiv:2004.10706v2. PMID: 32510522; PMCID: PMC7251955.

Shweta Yadav, Srivastsa Ramesh, Sriparna Saha, and Asif Ekbal. 2020. Relation extraction from biomedical and clinical text: Unified multitask learning framework. *IEEE/ACM Trans Comput Biol Bioinform.*

Shi Yuan and Bei Yu. 2018. An evaluation of information extraction tools for identifying health claims in news headlines. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 34–43. Association for Computational Linguistics.

Rui Zhang, Dimitar Hristovski, Dalton Schutte, Andrej Kastrin, Marcelo Fiszman, and Halil Kilicoglu. 2021. Drug repurposing for COVID-19 via knowledge graph completion. *Journal of biomedical informatics*, 115(103696).