Automatic recognition of figurative language in biomedical articles

Dina Demner-Fushman, Willie Rogers, James Mork
National Library of Medicine
8600 Rockville Pike
Bethesda, MD, 20894
{ddemner,wjrogers,jmork}@mail.nih.gov

Abstract

Figurative language plays an important role in thought processes and science. Automatic detection of figurative language is gaining momentum in the open domain natural language processing research, but it is hindered in the biomedical domain by the absence of document collections for development and testing of the approaches. Reliable approaches to detection of figurative language could potentially improve automatic indexing of the literature and support clinical applications. We have developed a collection of documents annotated for literal or non-literal use of seven terms that are known to cause errors in automatic indexing of biomedical abstracts. Using the collection, we explore detection of figurative language with CNN-RNN, logistic regression and transformer models. We establish baselines for each of the seven terms, achieving the results at the level of the state-of-the-art reported in the open domain evaluations.

Introduction

Figurative language plays an important role in science, with metaphors and idiomatic expressions viewed as foundations for thought processes (Taylor and Dewsbury 2018; Cork, Kaiser, and White 2019). Wide use of figurative language in the biomedical literature presents a significant challenge in automatic text understanding. Consider the term *falls* in the following sentences:

A patient who suffered a fall from a wagon.

Falling off the care wagon.

Falling off the dopamine wagon.

Fall from a train wagon.

Fall from horse-drawn wagon.

Whereas it is relatively easy for people to discern which of these phrases refer to physical falls, the biomedical named entity recognition (NER) approaches often treat figurative language as literal and link the word to inappropriate ontology terms as a result. Specifically, in the task of automated indexing that aims to summarize the main points of a publication by assigning terms from a controlled vocabulary created to index the biomedical literature: Medical Subject Headings (MeSH) (NLM 2020 (accessed November, 2020).

No copyright. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In the biomedical publications, the problem of recognizing non-literal utterances is intertwined with word sense disambiguation (WSD), and compounded by the importance of the term to the article. The WSD aspects could be illustrated by the following:

> The head of each fish, including the brain and pituitary, was sampled for double-colored FISH analysis.

To many NER approaches, the first occurrence of fish is indistinguishable from FISH, which stands for fluorescent in situ hybridization. The confusion continues in:

> Is being a small fish in a big pond bad for students´ psychosomatic health?

Moreover, for food products manufactured from fish, such as fish oil, linking to *Fishes* also violates indexing rules. To summarize, to label a biomedical publication with the terms from a terminology, we need to determine if the terms are used literally, if the sense in the context corresponds to the sense in the terminology, and if the term is important enough to be indexed for the article in MEDLINE/PubMed database, which comprises more than 30 million biomedical abstracts (NLM 2020 (accessed November, 2020). The importance of a term plays a bigger role when we use the existing manual indexing of biomedical abstracts for training and testing: The correct sense of a term could be used literally in the abstract, but the term might not be central enough to the publication to be assigned by the indexer.

Whereas there continues to be a steady research in biomedical WSD (Pesaranghader et al. 2019), and use of figurative language in biomedicine (Cork, Kaiser, and White 2019), automated understanding of biomedical figurative language is still an under-explored area. Our objectives therefore are:

1. to determine which non-literal expressions are prevalent in the biomedical literature and present difficulties to automated understanding,
2. create training and test collections for these terms, and
3. explore approaches to automated detection of non-literal language.

Related Work

The body of work on detection of figurative language in the open domain is significant, and the interest to the topic is

CEUR-WS.org/Vol-2831/paper9.pdf

growing, as evidenced by the workshops and shared tasks on figurative language processing (Klebanov et al. 2020). Veale et al. (2016) provide an overview of the types of figurative language and of the computational approaches to detection and understanding of figurative language. The approaches are mostly formulated as a binary classification task on a limited set of triples, and sometimes as prediction of the class of a token in a sentence (Feldman and Peng 2013; Gao et al. 2018). Taking into account the immediate lexico–syntactic context of the utterance and incorporating discourse features improves recognition of figurative language (Mu, Yannakoudakis, and Shutova 2019). In an end-to-end RNN-based system, Mao et al. (2019) emulated two human approaches to identification of figurative language: 1) noticing a semantic contrast between a target word and its context – *Selectional Preference Violation*, and 2) identifying if the literal meaning of a word contrasts with the meaning that word takes in the context – *Metaphor Identification Procedure*.

To the best of our knowledge, our work is the first to explore the difficulties figurative language poses for automated indexing of the biomedical literature. We also provide the first publicly available biomedical literature dataset annotated for figurative language at the token and sentence level. In addition, leveraging the state-of-the-art approaches explored in the open domain, we establish baselines for detection of figurative language in biomedical abstracts using sentence or token level classification.

Data Sources and Collections

We analyzed 870 American English idioms (Bulkes and Tanner 2017), and 464 metaphors (Katz et al. 1988; Campbell and Raney 2016). We searched the Free Dictionary Idioms dictionary (FARLEX 2020 (accessed November, 2020) for additional examples of figurative phrases. We then submitted figurative language expressions to MeSH on Demand (NLM 2020 (accessed November, 2020) to identify potential triggers for false-positive linking to MeSH e.g., *cat* and *mouse* in "the game of cat and mouse" could be mapped to *Cats* and *Mice*, respectively. We then searched PubMed with these trigger terms to get the frequency of their use in publications. We identified seven most frequent false positives triggers that are shown in Table 1 along with the sizes of the training and test sets for each term.

We then searched PubMed for the exact figurative expressions, and for the abstracts containing trigger terms that were either indexed or not with the corresponding MeSH headings. Abstracts with trigger terms and MeSH headings serve as examples of literal use in the training set, and abstracts without MeSH headings serve as examples of non-literal use. For the test sets, we randomly sampled files from both distributions and manually annotated the sentences containing the terms at the token level. We annotated fine-grained senses corresponding to:

1. **Full MH**: the literal Mesh Heading-appropriate sense, e.g., "a healthy baby at 34 weeks of gestation." The labels assigned by the indexers were not shown to the annotators to avoid bias.

Term (MH)	Check Tag	Training	Test
fall (Accidental Falls)	no	45,820	895
fish (Fishes)	no	18,256	513
juvenile (Adolescent)	yes	59,176	581
baby (Infant)	yes	1,065	270
bull (Cattle)	yes	1,194	555
cat (Cats)	yes	4,368	542
dog (Dogs)	yes	19,167	905

Table 1: Sizes of the training and test sets for each term in the PubMed Figuratively Language Collection. The Check Tag column indicates if the term is a required term to be added because it pertains to the subject of the study. Check Tags are the most frequently used MeSH terms, which indicates our collection covers a sizable portion of false positive triggers.

2. **Partial Literal**: MH-appropriate sense, but being a part of an expression, which should not trigger mapping to MeSH, e.g., *shaken baby syndrome*.
3. **Literal Other**: Literal senses other than MH, e.g., baby hamster is still a baby, but it should not be indexed with *Infant*, which applies only to human babies.
4. **Figurative**: Non-literal use of the term, e.g., in "There's a Baby in this Bath Water!"

Each document was annotated by two annotators and the differences were reconciled.

Experiments

We explored CNN-RNN (Svoboda 2020 (accessed November, 2020), Logistic Regression (Pedregosa et al. 2011) and BERT-based (Kaiyinzhou 2020 (accessed November, 2020) approaches with various embeddings and the Universal Sentence Encoder (Cer et al. 2018). We used sentences from PubMed abstracts containing the trigger terms and the expressions from the above collections of idioms for training these models. Due to sparseness of the annotations and unavailability of sufficient examples for training and for judging the results, we collapsed the annotations into two classes: figurative or literal MH-appropriate. Any terms that were labeled *LiteralOther* or *PartialLiteral* were relabeled as *Figurative*. For example, in an article about dog owners, dog was considered as non-literal. Terms labeled as *Figurative* or *FullMH* remained unchanged.

We then approached the task as binary classification at the sentence or token level.

To train the CNN-RNN and Logistic Regression models, sentences containing the target trigger terms were extracted from a set of retrieved documents that were labeled using MeSH indexing information as described above. Each extracted sentence was assigned the label of the document from which it was derived. Sentence embeddings were generated using a Doc2Vec (Rehurek and Sojka 2010) model pre-trained on the documents retrieved for the trigger terms.

In the CNN-RNN approach, the embeddings and associated labels served as input to a neural network containing four groups of four layers: convolutional layer, dropout, max-pooling, and dropout, followed by an LSTM layer.

Term	Sentence level CNN-RNN				Sentence level Logistic regression				Sentence level USE				Token level BERT			
	P	R	F1	A	P	R	F1	A	P	R	F1	A	P	R	F1	A
fall	0.77	0.68	0.72	**0.99**	0.64	0.78	0.71	0.73	**0.89**	**0.89**	**0.89**	0.88	0.37	0.34	0.35	0.98
fish	0.51	0.48	0.50	**0.99**	**0.58**	0.45	0.50	0.50	**0.58**	**0.54**	**0.56**	0.48	0.37	0.35	0.36	0.98
juvenile	0.77	0.64	0.70	**0.99**	**0.97**	0.38	0.55	0.86	0.82	**0.83**	**0.82**	0.80	0.37	0.36	0.37	0.99
baby	**0.76**	**0.99**	**0.86**	0.99	0.39	0.36	0.37	0.39	0.67	0.56	0.61	0.45	0.61	0.61	0.61	**0.99**
bull	**0.90**	**0.87**	**0.88**	0.99	0.56	0.38	0.45	0.58	0.78	0.74	0.76	0.71	0.84	0.86	0.85	**0.99**
cat	**0.77**	0.74	**0.76**	0.99	0.54	0.74	0.63	0.54.	0.73	0.73	0.73	0.65	0.68	**0.78**	0.73	**0.99**
dog	**0.76**	**0.97**	**0.85**	0.98	0.48	0.55	0.51	0.50	0.63	0.58	0.60	0.65	**0.76**	0.78	0.77	**0.99**

Table 2: Results of predicting literal and figurative use of trigger terms. USE = Universal sentence encoder, R = Recall, P = Precision A = Accuracy. The differences in 0.99 accuracy between the CNN-RNN and BERT approaches are in the third decimal point.

The model uses a sigmoid activation function, binary cross-entropy loss and the adam optimizer.

We used the SciKit Learn Logistic Regression classifier, with Doc2Vec output as inputs.

The Universal Sentence Encoder was also applied in the sentence level classification task. Unlike the Doc2Vec models, the Universal Sentence Encoder was trained on a very large corpus using a variety of sources. In our approach, each sentence vector representation was generated using the Universal Sentence Encoder during training. The vector representation and the sentence label was then passed to a two-layer neural network consisting of a RELU and a softmax layer. A categorical cross-entropy loss and the adam optimizer was used when building the model.

We used BERT encoder extended with a CRF layer for Named Entity Recognition (Kaiyinzhou 2020 (accessed November, 2020) for the token-level classification of literal and figurative use of the tokens. We used BIO-style (Beginning-Inside-Outside) features. To train BERT, we tagged the trigger terms with the label of the sentence and all other terms in the sentence as outside.

Results

Table 2 summarizes the results obtained for the binary classification approaches to detection of figurative language. The PubMed searches yielded training sets of varying sizes, ranging from $1,065$ documents for *baby*, to $59,176$ for *juvenile*. The manually annotated test sets for each of the terms range from 270 to 905 documents. The size of the training set does not seem to be directly correlated with the results, as shown in Figure 1.

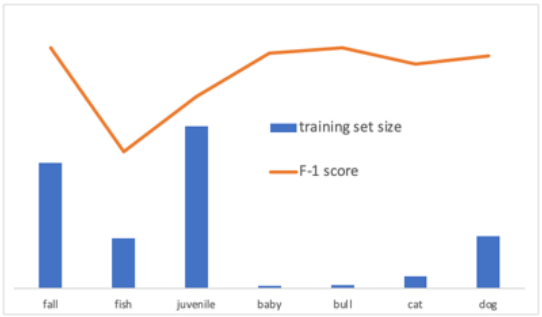

Figure 1: The size of the training set does not always directly influence the best F-1 scores obtained in figurative language detection

Discussion

We created a collection of PubMed abstracts automatically annotated for literal and non-literal use of seven terms that proved to be a rich source of false positive linking to terminologies and have sufficient amounts of training documents in PubMed. Interestingly, one of these terms, *fall* was also found to be difficult to classify as figurative in the open domain tasks (Stowe et al. 2019).

We explored several state-of-the-art approaches, casting the task as binary classification at the sentence and token level. We hoped to identify one best approach for the task and achieve state-of-the-art performance for all trigger terms. The best results reported in the literature for the open-domain figurative language detection and in the shared task on metaphor detection (Klebanov et al. 2020) are around 70% F-1 score, sometimes reaching 80% and above performance. Although we have obtained F-1 scores above 80% for five of the seven terms, we cannot identify a single approach that will achieve good scores on all trigger terms. The F-1 score for *fish* is only 56%. This score could probably be explained by the fact that this term often violates the widely used WSD assumption of "one sense per document" (Yarowsky 1995), which we used to create the training set. As can be seen in the example, two senses of *fish* are used in the same sentence:

> These preliminary results provide the basis for the further development of a non-GMO approach to modulate fish allergenicity and improve safety of aquaculture fish. (PMID: 31622806)

The indexers labeled this article with both *Fishes* and *Seafood*. When the contexts for these occurrences of *fish* are used in the models as positive examples, they might be too close to the contexts of the articles that present *fish* only in the context of food and thus serve as negative examples.

With respect to identifying one approach that would work

best for all of the trigger terms, we can see that casting the task as sentence-level classification and using the CNN-RNN model produces the majority of best results. Stowe (2019) observes that *fall* is difficult to classify because the distribution of the literal and metaphoric uses of this word in the open domain is almost even. In our annotations, we also observed frequent use of *fall* in personification, which might explain why the Universal Sentence Encoder pre-trained on a variety of sources performs much better for *falls*.

Another interesting observation is that if we want to select a method for automated indexing, we will have to decide if recall or precision are more important when suggesting the terms. For *cat*, *dog*, *fish* and *juvenile*, the differences in these two metrics achieved by different approaches are relatively large, although the F-scores are mostly close, showing a typical trade-off between the two metrics. In selecting approaches to support automated indexing, precision often plays an important role, as currently the consensus is that it is better to miss a term than to assign an inappropriate term that will mislead the search engines that rely on MeSH indexing. For that reason, we do not consider accuracy when selecting an approach for supporting automated indexing.

Our work has some limitations that we hope to address in the future. First, we addressed only seven of the hundreds of terms used figuratively in the biomedical literature. Although the seven terms provided enough information to see that no single approach is a winning strategy, additional annotations will be needed for testing approaches to figurative language detection on PubMed scale. We also found that for many remaining terms figurative use in PubMed is infrequent and additional sources of figurative language will be needed for training. For example, *butterflies in my stomach* is used in PubMed only two times, and *butterflies* AND *stomach* 20 times. More data will be needed to train a classifier to distinguish between these two titles:

> Butterflies in My Stomach: Insects in Human Nutrition
>
> Neurotic butterflies in my stomach: the role of anxiety, anxiety sensitivity and depression in functional gastrointestinal disorders

Conclusions

This work presents an initial exploration of the use and detection of figurative language in biomedical publications. On the one hand, figurative language is known to play an important role in thought processes and in science, and therefore being widely used in biomedical publications, on the other hand, automated detection of figurative language in the biomedical publications has not yet attracted research. To explore feasibility of automated detection of figurative language, we created a collection of documents annotated for literal or non-literal use of seven terms that are known to cause errors in automatic indexing of biomedical abstracts with MeSH terms. We then explored sentence and token-level classification approaches to detection of figurative language using CNN-RNN, logistic regression and transformer models. With the exception of one term, *fish*, our performance is on par with the state-of-the-art achieved in the open domain evaluations. We hope that the interesting problem of detection of figurative language in biomedical text, the dataset, and the automated approach to creation of the training sets outlined in this work will bring about further research in this area.

Data & code: https://ii.nlm.nih.gov/DataSets/index.shtml

Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

We thank Alan Aronson, Francois Lang, Laritza Rodriguez and Sonya Shooshan for judging parts of the collections. We thank Anna Ripple for constructing PubMed searches.

References

Bulkes, N. Z.; and Tanner, D. 2017. "Going to town": Large-scale norming and statistical analysis of 870 American English idioms. *Behavior research methods* 49(2): 772–783.

Campbell, S. J.; and Raney, G. E. 2016. A 25-year replication of Katz et al.'s (1988) metaphor norms. *Behavior research methods* 48(1): 330–340.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169–174.

Cork, C.; Kaiser, B. N.; and White, R. G. 2019. The integration of idioms of distress into mental health assessments and interventions: a systematic review. *Global Mental Health* 6.

FARLEX. 2020 (accessed November, 2020). *25. The Free Dictionary by FARLEX. Idioms and phrases.* URL https://idioms.thefreedictionary.com/.

Feldman, A.; and Peng, J. 2013. Automatic detection of idiomatic clauses. In *International Conference on Intelligent Text Processing and Computational Linguistics*, 435–446. Springer.

Gao, G.; Choi, E.; Choi, Y.; and Zettlemoyer, L. 2018. Neural Metaphor Detection in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 607–613.

Kaiyinzhou. 2020 (accessed November, 2020). *BERT-NER*. URL https://github.com/kyzhouhzau/BERT-NER.

Katz, A. N.; Paivio, A.; Marschark, M.; and Clark, J. M. 1988. Norms for 204 literary and 260 nonliterary metaphors on 10 psychological dimensions. *Metaphor and Symbol* 3(4): 191–214.

Klebanov, B. B.; Shutova, E.; Lichtenstein, P.; Muresan, S.; Wee, C.; Feldman, A.; and Ghosh, D., eds. 2020. *Proceedings of the Second Workshop on Figurative Language Processing*. Online: Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.figlang-1.0.

Mao, R.; Lin, C.; and Guerin, F. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3888–3898.

Mu, J.; Yannakoudakis, H.; and Shutova, E. 2019. Learning Outside the Box: Discourse-level Features Improve Metaphor Identification. In *Proceedings of NAACL-HLT*, 596–601.

NLM. 2020 (accessed November, 2020)a. *Medical Subject Headings*. URL https://www.nlm.nih.gov/mesh/meshhome.html.

NLM. 2020 (accessed November, 2020)b. *MEDLINE and PubMed*. URL https://pubmed.ncbi.nlm.nih.gov/.

NLM. 2020 (accessed November, 2020)c. *MeSH on Demand*. URL https://meshb.nlm.nih.gov/MeSHonDemand.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Pesaranghader, A.; Matwin, S.; Sokolova, M.; and Pesaranghader, A. 2019. deepBioWSD: effective deep neural word sense disambiguation of biomedical text data. *Journal of the American Medical Informatics Association* 26(5): 438–446.

Rehurek, R.; and Sojka, P. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Stowe, K.; Moeller, S.; Michaelis, L.; and Palmer, M. 2019. Linguistic Analysis Improves Neural Metaphor Detection. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 362–371.

Svoboda, D. 2020 (accessed November, 2020). *Doc2Vec*$_C NN_R NN.URL$.

Taylor, C.; and Dewsbury, B. M. 2018. On the problem and promise of metaphor use in science and science communication. *Journal of microbiology & biology education* 19(1).

Veale, T.; Shutova, E.; and Klebanov, B. B. 2016. Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies* 9(1): 1–160.

Yarowsky, D. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, 189–196.