MetaMap versus BERT models with explainable active learning: ontology-based experiments with prior knowledge for COVID-19

M. Arguello-Casteleiro[1], C. Henson[2], N. Maroto[3], S. Li[4], J. Des-Diz[5], M.J. Fernandez-Prieto[6], S. Peters[1], T. Furmston[1], C. Sevillano Torrado[5], D. Maseda Fernandez[2], M. Kulshrestha[2], J. Keane[1], R. Stevens[1], and C. Wroe[7]

[1] University of Manchester, UK
[2] Midcheshire Hospital Foundation Trust, UK
[3] Universidad Politécnica de Madrid, Spain
[4] University of Stirling, UK
[5] Hospital do Salnés, Spain
[6] University of Salford, UK
[7] BMJ, UK

Abstract. Emergence of the Coronavirus 2019 Disease has highlighted further the need for timely support for clinicians as they manage severely ill patients. We combine Semantic Web technologies with Deep Learning for Natural Language Processing with the aim of converting human-readable best evidence/practice for COVID-19 into that which is computer-interpretable. We present the results of experiments with 1212 clinical ideas (medical terms and expressions) from two UK national healthcare services specialty guides for COVID-19 and three versions of two BMJ Best Practice documents for COVID-19. The paper seeks to recognise and categorise clinical ideas, performing a Named Entity Recognition (NER) task, with an ontology providing extra terms as context and describing the intended meaning of categories understandable by clinicians. The paper investigates: 1) the performance of classical NER using MetaMap versus NER with fine-tuned BERT models; 2) the integration of both NER approaches using a lightweight ontology developed in close collaboration with senior doctors; and 3) the easy interpretation by junior doctors of the main classes from the ontology once populated with NER results. We report the NER performance and the observed agreement for human audits.

Keywords: Ontologies, Deep Learning for Natural Language Processing, static embeddings, transformer-based language models, COVID-19

1 Introduction

The World Health Organization declared the Coronavirus Disease 2019 (COVID-19) outbreak a pandemic on 11 March 2020 [1]. The COVID-19 pandemic is a prime example of a need for evidence-based recommendations for clinical care that need to be reviewed and updated frequently and rapidly. There are datasets for COVID-19 such as CORD-19 [2] and LitCovid [3] that are updated regularly, including thousands of articles from PubMed/MEDLINE [4]. However, not all the publications included in those datasets have the same clinical value as resources for Evidence-Based Medicine (EBM) [5] that integrates clinical experience with the best scientifically sound research available [5].

The body of scientific evidence for healthcare is not limited to information in PubMed/MEDLINE articles, but also includes clinical point-of-care summaries and clinical practice guidelines from healthcare services. BMJ Best Practice [6] and UpToDate [7] are examples of clinical evidence summaries that aim to bring the latest evidence from health research into healthcare practice.

Biomedical facts and clinical recommendations are made of natural language statements, typically complex sentences that are human-readable and intended for expert-to-expert communication. Named Entity Recognition (NER) is one well-known natural language processing (NLP) task that seeks to recognise specific words or phrases ('entities') from natural language statements and categorise them [8]. In this study, we adhere to a functional perspective on ontologies [9], and explore how ontologies can be used for categorisation in support of NER task. We consider ontologies as artifacts that can [9]: a) provide background knowledge about a domain; b) contain a list of terms associated with the ontology's classes and relations; and c) supply formal machine-readable definitions and axioms represented in many forms.

This paper addresses three critical questions regarding prior knowledge for COVID-19, i.e. best evidence/practice provided by UK clinical practice guidelines and BMJ Best Practice documents for COVID-19. Firstly, what is the performance of classical NER using MetaMap [10] versus the state-of-the-art NER with transformer-based language models from Deep Learning for NLP [11], such as BERT (Bidirectional Encoder Representations from Transformers) [12]? Secondly, to what extent does a lightweight ontology facilitate the integration of results from both NER approaches? Thirdly, to what extent can the main classes from the lightweight ontology, once populated with NER results, be easily interpreted by junior doctors?.

The novelty of this paper is three-fold: 1) presenting the Evidence-Based Recommendation Ontology (EBRO), a light-weight ontology co-created with close collaboration with senior doctors (medical consultants from UK and Spain) that aims to contain main classes easily interpretable by junior doctors; 2) proposing a different problem formulation for NER as a fine-tuning specific task with transformer-based language models, considering NER as a sequence-level task instead of a token-level task [12]; and 3) exploring NER performance for BMJ Best Practice text excerpts for COVID-19 using biomedical-specific transformer models and general-domain transformer models (e.g. BERT) fine-tuned for NER with titles and available abstracts from PubMed/MEDLINE articles about COVID-19.

The approach presented follows Semantic Deep Learning [13] combining Semantic Web technologies and Deep Learning for NLP. The paper belongs to explainable artificial intelligence [14]. The fine-tuning of transformer-based language models fits in the new field of explainable active learning (XAL) [15], differing from traditional active learning (AL) in providing the model's prediction together with an explanation.

2 Experiments with prior knowledge for COVID-19

We start by presenting the informal and formal meanings for the main classes within the EBRO. Next, we illustrate how the outcome of both MetaMap and transformer-based language models fine-tuned for NER can be incorporated into the EBRO. We introduced six principles developed to assess the outcome from MetaMap. We provide details of XAL setup for fine-tuning transformer-based language models for NER. Finally, we recapitulate the experimental design and the measures for evaluating the performance of the experiments, including human audits.

2.1 The Evidence-Based Recommendation Ontology (EBRO)

This study presents the EBRO, an ontology represented in the W3C Web Ontology Language (OWL) [16]. We take a pragmatic approach to the ontology building and prioritise re-use over other considerations. The EBRO reuses axioms from several ontologies, such as: the Ontology Lexicon (Ontolex) [17]; the Semantic science Integrated Ontology (SIO) [18]; the Basic Formal Ontology (BFO) [19] and the Information Artifact Ontology (IAO) [20]. The EBRO can be downloaded from [21].

Table 1 illustrates informal descriptions of the main EBRO classes according to senior doctors along with the more formal descriptions in the Manchester OWL syntax [22] with classes from BFO and IAO. The EBRO reuses some of the Unified Medical Language System (UMLS) Semantic Types [23] like ‘T121|Pharmacologic Substance’. The EBRO reuses the classes ‘Condition’ and ‘Population’ from the PICO ontology [24].

Table 1. Illustrating informal descriptions for EBRO classes

EBRO class	Informal description	Manchester OWL syntax
Patient's healthcare problem	A healthcare problem implies the presence of clinical findings including symptoms, normal/abnormal clinical states, and diagnoses.	SubClassOf: 'obo:realizable entity'
Process of care	“The processes through which patient care is delivered” [25].	SubClassOf: obo:process
Patient’s treatment	“Action taken by a health professional, in the context of contact with a treatment recipient, to alter the functioning of an individual with a disability or at risk of a disability” [26].	SubClassOf: 'Process of care'
Patient’s test	"All types of tests are eligible” [27].	SubClassOf: 'Process of care'
Chemicals & Drugs	Some UMLS Semantic Types like ‘Pharmacologic Substance’ are included as subtypes.	SubClassOf: 'obo:material entity'
Evidence-based information source	Examples are: PubMed articles; clinical evidence summaries (e.g. BMJ Best Practice); and clinical practice guidelines.	SubClassOf: obo:document and ('obo:has evidence' some obo:evidence)

2.2 EBRO and NER: MetaMap versus BERT models

Figure 1 sketches how EBRO incorporates the NER results of both: (1) classic NER with MetaMap; and (2) NER with fine-tuned transformer-based language models.

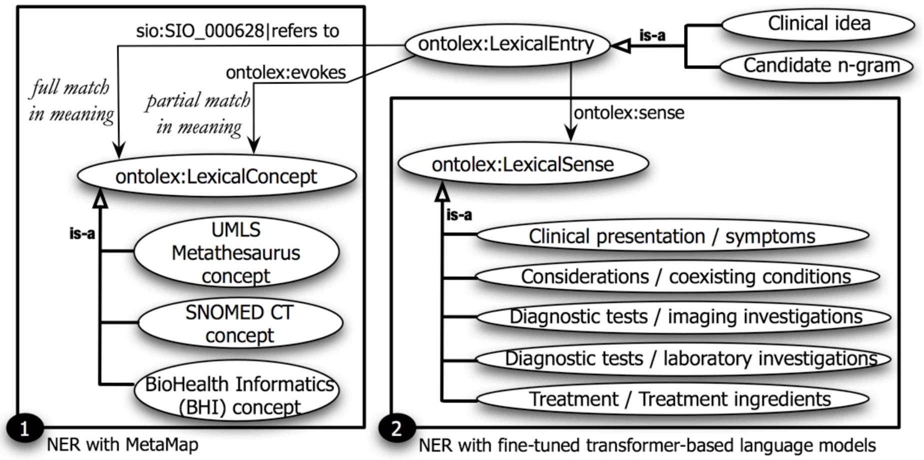
sio:SIO_000628|refers to
ontolex:LexicalEntry
is-a
Clinical idea
Candidate n-gram
ontolex:evokes
full match in meaning
partial match in meaning
ontolex:sense
ontolex:LexicalConcept
ontolex:LexicalSense
is-a
UMLS Metathesaurus concept
SNOMED CT concept
BioHealth Informatics (BHI) concept
is-a
Clinical presentation / symptoms
Considerations / coexisting conditions
Diagnostic tests / imaging investigations
Diagnostic tests / laboratory investigations
Treatment / Treatment ingredients
1
NER with MetaMap
2
NER with fine-tuned transformer-based language models

Fig. 1. Overview of how to incorporate NER results into the EBRO

NER with MetaMap. To express a clinical idea, a medical term - more generally, a medical expression - may be needed. A clinical idea may match fully or partially a concept from existing clinical/biomedical terminologies. In UMLS Metathesaurus [10], each concept has a Concept Unique Identifier (CUI) and one or more UMLS Semantic Types. In UMLS, a Metathesaurus concept is mapped to zero, one, or more than one concept from the clinical terminology SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) [28]. MetaMap can map a clinical idea to UMLS Metathesaurus concepts and SNOMED CT concepts. Occasionally, professional terminologists may suggest a bio-health informatics (BHI) concept that merges multiple similar UMLS CUIs into one concept or is a concept non-existent in UMLS.

Six principles to assess NER with MetaMap. This study selects the focus concept(s) among the CUIs provided by MetaMap, but in some cases the mapping has been performed manually. The selection of the focus concept(s) is guided by six principles:

1. The focus concept is interpreted in this study as the CUI that captures the key and more specific biomedical/clinical meaning (i.e. governing term).
2. When selecting the focus concept, avoid general biomedical/clinical terms in favour of more specific terms.
3. When selecting the focus concept, favour CUIs that have a wider coverage in vocabulary sources as well as a wider meaning, and if pertinent, are already included in SNOMED CT. If the CUI covers "literally" the clinical idea, this should be selected, even if it is Not in SNOMED CT. If the CUI covers the clinical idea and is mapped to SNOMED CT, this should be selected.

4. A focus concept can have one or more refinements (i.e. dependent terms).
5. Negation is interpreted in this study as a refinement.
6. Multiple focus concepts should be considered only if there is more than one governing term, and if possible, belonging to the same UMLS Semantic Type. When using multiple focus concepts the meaning is their combined meaning, i.e. with logical "OR" for connecting the multiple focus concepts.

NER as fine-tuning transformer-based language models. In this work, we use the BERT base model as a baseline [12]. For the experiments, we consider the general-domain pre-trained language models BERT and RoBERTa [29]. We also consider five biomedical-specific pre-trained language models: BioBERT [30], SciBERT [31], ClinicalBERT [32], BlueBERT [33], and PubMedBERT [11]. Table 2 has information about the corpus from which transformer-based models were pre-trained. SciBERT was pre-trained using PubMed Central [34] and computer science (CS) literature. The model names (last column) are from the python library transformers by Hugging Face [35], which is used to fine-tune the transformer-based models.

The fine-tuning of BERT for a NLP downstream task is a problem formulation with two alternatives: token classification or sequence classification [11,12]. The problem formulation for NER is token classification [11,12]. A novelty of our work is to consider NER as sequence classification like question answering, and thus, the AutoModelForSequenceClassification implemented in Hugging Face [35] is utilised.

Table 2. BERT models: information about the pre-trained transformer-based models.

Pre-trained model	Corpus	Hugging Face transformers [35]: model name
BERT	Wiki + Books	bert-base-cased
RoBERTa	Web crawl	roberta-base
BioBERT	PubMed	dmis-lab/biobert-base-cased-v1.1
SciBERT	PubMed Central + CS	allenai/scibert_scivocab_cased
ClinicalBERT	MIMIC	emilyalsentzer/Bio_ClinicalBERT
BlueBERT	PubMed + MIMIC	bionlp/bluebert_pubmed_uncased_L-24_H-1024_A-16
PubMedBERT	PubMed	microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract

NER fine-tuning with XAL setup. Figure 2 outlines the AL cycle. We used word2phrase from word2vec [36] to obtain n-grams for each PubMed dataset (fourth column in Table 3). Each PubMed article has a unique identifier (PMID). The experimental setup for XAL considers 10 iterations, leveraging on word2vec models [36] created with the skip-gram algorithm using titles and available abstracts from PubMed/MEDLINE articles. For each iteration, the model Mi from Table 3 with i=[1,10] provided some instances for training the transformer-based language models.

M1 to M6 are created with titles and available abstracts (raw text) from PMIDs that appear among the bibliographic references of the BMJ Best Practice for COVID-19

[37] released around the date shown in Table 3 (first column). M7 to M10 are created with titles and available abstracts from files downloaded from PubMed from December 2019 until the date displayed in Table 3 and having terms such as 'COVID-19', 'SARS-CoV-2', and 'coronaviruses' in the title, abstract, and original subject headings. The last column in Table 3 has the number of PMIDs from CORD-19 [2] dataset of 31 May 2021. Comparing the last two columns of Table 3, few PMIs included in the BMJ Best Practice for COVID-19 [37] are not included in CORD-19.

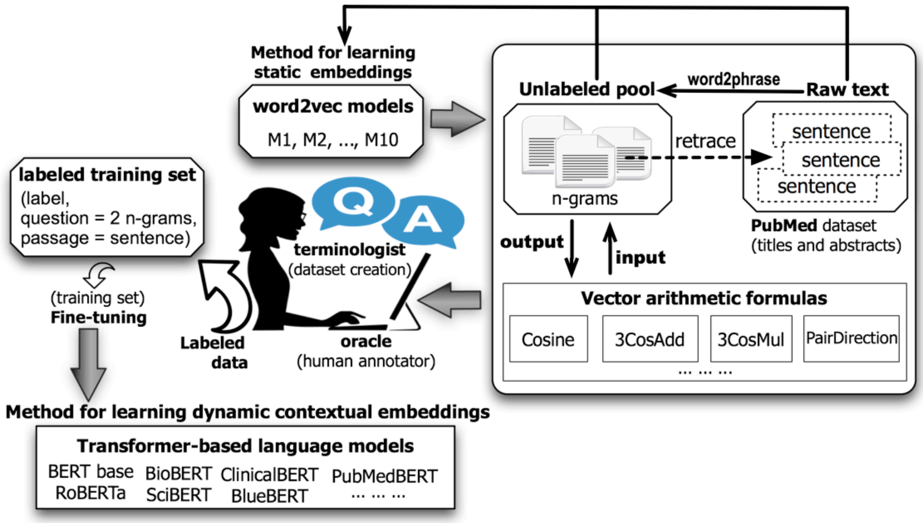

Fig. 2. Overview of NER fine-tuning for BERT models with XAL setup

Table 3. Information about the word2vec models created and used for XAL

Date	word2vec model	word2vec algorithm	PubMed dataset (number of PMIDs)	Number of PMIDs in CORD-19
10-August-2020	M1	skip-gram	706	703
10-July-2020	M2, M3, M4	skip-gram	671	668
10-June-2020	M5	skip-gram	594	593
11-May-2020	M6	skip-gram	435	434
10-August-2020	M7	skip-gram	41,472	40,876
10-July-2020	M8	skip-gram	32,245	31,809
10-June-2020	M9	skip-gram	22,513	22,216
11-May-2020	M10	skip-gram	11,771	11,670

The fine-tuning can be construed as reading comprehension [38] with a training dataset having 3-tuples (label, question, passage). The label is the prediction True/False also interpretable as a yes/no answer for question answering task. The question consists of an n-gram representing a general medical term — providing the explicit meaning and conveying a lexical sense (see Figure 1) — together with an input/output n-gram (appearing as 'candidate n-gram' in Figure 1) from vector arithmetic formulas [39] applied to word2vec model Mi. The passage is a sentence obtained by retracing the input/output n-gram into the PubMed dataset, which was re-organised by date and source. The sentence acts as a local explanation, justifying only the reason for the prediction on a specific input instance [40].

The vector arithmetic formulas [39] (see Figure 2) act as the active learning sampling strategy, i.e. the scoring functions to select the 'candidate n-gram' for the queries to fine-tune pre-trained transformer-based language models. The total number of unique n-grams from models M1 to M10 is 98,901 and the vector arithmetic formulas selected 2060 instances for training: 1575 are False and 485 are True.

2.3 Experimental design and evaluation metrics

Figure 3 distills the essence of the experiments conducted and the human audits performed involving two professional terminologists and one junior doctor.

NER experiments with prior knowledge for COVID-19. For NER with MetaMap we used 1212 clinical ideas (see 'clinical idea' in Figure 1) appearing in textual excerpts from two UK national healthcare service specialty guides for COVID-19 [41] and two BMJ Best Practice documents for COVID-19 [37,42]. We considered a 3-month chronology: documents released around 10th of May, June, and July 2020. For NER with fine-tuned transformer-based models we used 259 clinical ideas appearing in 345 textual excerpts (interpreted here as passage) that are new in the June 2020 version when compared with May 2020 version of the two BMJ Best Practice documents for COVID-19 [37,42].

NER performance metrics using the human gold standard labels. A professional terminologist and a senior doctor, both with many years of experience as clinical coders, provide the human gold standard labels [38]. We report precision, recall, and F-measure [38] for NER using the human gold standard labels.

For NER with MetaMap, considering the six principles introduced in the previous subsection, there are two possibilities when mapping the meaning of a clinical idea to UMLS CUI(s): a) full match in meaning, e.g. synonym, expressed as a "*a clinical idea is-a focus concept, which 'refers to (full match)' CUI*"; b) partial match in meaning, i.e. something is not captured by the CUI(s), expressed as "*a clinical idea has at least one focus concept, which 'evokes (partial match)' CUI*". If there are multiple focus concepts, each single CUI is a partial match in meaning, i.e. "evokes".

For NER with fine-tuned transformer-based language models, a clinical idea may appear in one or more textual excerpts from BMJ Best Practice. Each textual excerpt is considered a passage. Every lexical sense from Figure 1 is systematically considered, i.e. composing questions with the clinical idea and the n-gram representing a general medical term. Transformers map sequences of input vectors {x1, ..., xn} to sequences of output vectors {y1, ..., yn} of the same length [38]. The NER result is interpreted by looking at the output label included in the output vector. The output label indicates if the clinical idea belongs (True/False) to the lexical sense.

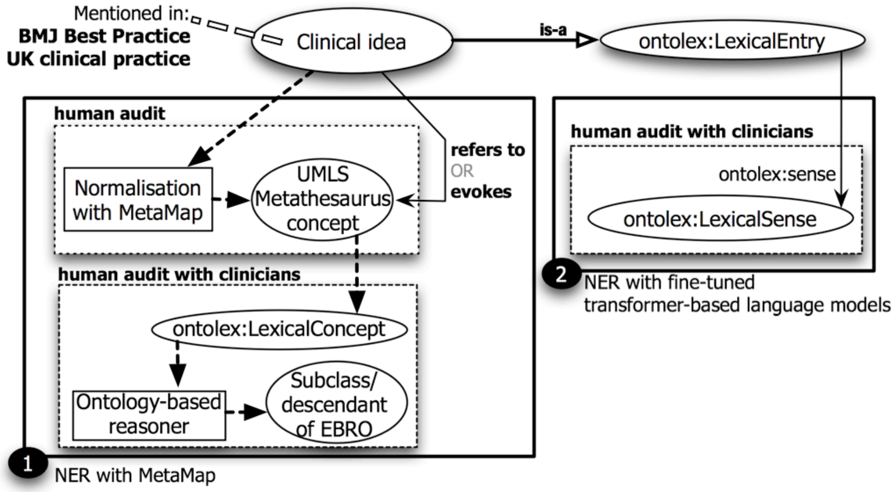

Fig. 3. Overview of the human audits in the experiments conducted for NER

Human audit: measuring agreement with the human gold standard labels. We carried out three human audits with domain experts as indicated in Figure 3. We report the observed agreement and kappa coefficient [43]. The human audit with clinicians judge the classification of the clinical ideas as: (1) children or descendants of EBRO main classes using the HermiT reasoner [44]; and (2) lexical senses (see Figure 1) relatable to the output labels included in the output vectors from BERT models.

3 Results: NER performance and human audit

Table 4 shows the performance for NER with BERT models for 259 clinical ideas.

Table 4. Performance for NER with excerpts and clinical ideas from BMJ Best Practice documents using BERT models fine-tuned for NER with COVID-19 articles from PubMed

Fine-tuned model	F1-measure %	Precision %	Recall %
BERT	78.05	67.69	92.15
RoBERTa	25.27	14.64	92.16
BioBERT	90.08	83.49	97.81
SciBERT	92.87	88.70	97.45
ClinicalBERT	82.14	73.12	93.70
BlueBERT	72.82	58.84	95.52
PubMedBERT	88.82	81.46	97.64

For the 1212 clinical ideas, NER with MetaMap using the six principles introduced earlier obtained F1-measure=90.96% with Precision=84.25% and Recall=98.83% for UMLS version 2016AA. For UMLS version 2020AA, F1-measure=91.93% with Precision=85.49% and Recall=99.42%.

Terminologist A had an observed agreement of 98.10% for 888 clinical ideas. Terminologist B had an observed agreement of 97.49% for 314 clinical ideas. Kappa K=0.887 for terminologists A and B is interpreted as "almost perfect agreement" [43].

For 972 children or descendants of EBRO main classes, a UK junior doctor had an observed agreement of 95.27%. For the 259 clinical ideas classified according to lexical senses, the same UK junior doctor had an observed agreement of 87.64%.

4 Concluding remarks

Whether physicians are ready to use evidence from big data remains unclear. However, this study suggests a high level of agreement by junior doctors for categories proposed by senior doctors. From an ontological point of view, the EBRO has many weaknesses, like including ambiguous lexical senses exploitable by BERT models. Indeed, the observed agreement is lower for the clinical senses after being populated for COVID-19 than the observed agreement for the EBRO main classes after being populated for COVID-19.

The six principles introduced to assess NER with MetaMap seem to foster a high agreement with professional terminologists and a performance quite close to NER with BERT models, where SciBERT obtained the highest F1-measure=92.87%.

References

1. WHO: COVID-19, https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020
2. CORD-19, https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
3. LitCovid, https://www.ncbi.nlm.nih.gov/research/coronavirus/
4. PubMed/MEDLINE, https://pubmed.ncbi.nlm.nih.gov/
5. Masic, I., Miokovic, M., Muhamedagic, B.: Evidence based medicine - new approaches and challenges. Acta Inform Med. 16(4), pp. 219-25 (2008).
6. BMJ Best Practice, https://bestpractice.bmj.com/info/
7. UpToDate, https://www.uptodate.com/home
8. Nadkarni, P.M., Ohno-Machado, L. and Chapman, W.W.: Natural language processing: an introduction. J Am Med Inform Assoc, 18(5), pp. 544-551 (2011).
9. Hoehndorf, R., Schofield, P.N., Gkoutos, G.V.: The role of ontologies in biological and biomedical research: a functional perspective. Brief Bioinform. 16(6), pp. 1069-80 (2015).
10. Aronson, A.R., Lang, F.: An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 17(3), pp. 229-236 (2010).
11. Gu, Y., Tinn, et al.: Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. doi: 10.1145/3458754 (2021).
12. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: 2019 NAACL, pp. 4171–4186 (2019).

13. Semantic Deep Learning, http://www.semantic-web-journal.net/content/special-issue-semantic-deep-learning
14. Gunning, D., et al.: XAI—Explainable artificial intelligence. Science Robotics. doi: 10.1126/scirobotics.aay712 (2019).
15. Ghai, B., et al. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. In: ACM proceedings on Human-Computer Interaction, pp. 1-28 (2021).
16. OWL, https://www.w3.org/TR/owl2-quick-reference/
17. Ontolex, https://www.w3.org/community/ontolex/wiki/Final_Model_Specification
18. SIO, https://bioportal.bioontology.org/ontologies/SIO
19. BFO, http://www.obofoundry.org/ontology/bfo.html
20. IAO, http://www.obofoundry.org/ontology/iao.html
21. EBRO, https://github.com/arguellocasteleiro/OWL/
22. Manchester OWL syntax, https://www.w3.org/TR/owl2-manchester-syntax/
23. UMLS Semantic Types, https://lhncbc.nlm.nih.gov/semanticnetwork/
24. PICO ontology, https://linkeddata.cochrane.org/pico-ontology
25. Medicare: A Strategy for Quality Assurance. doi: 10.17226/1547 (1990).
26. Hart, T., et al.: Toward a theory-driven classification of rehabilitation treatments. doi: 10.1016/j.apmr.2013.05.032 (2014).
27. Cochrane Handbook for Systematic Reviews of Interventions, Version 6.2, 2021. https://training.cochrane.org/handbook/current/chapter-i
28. SNOMED CT, https://www.snomed.org/snomed-ct/five-step-briefing
29. Liu, Y., et al.: Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692 (2019).
30. Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), pp.1234-1240 (2020).
31. Beltagy, I., Lo, K. and Cohan, A.: Scibert: A pretrained language model for scientific text. In: 2019 EMNLP-IJCNLP, pp. 3615–3620 (2019).
32. Alsentzer, E., et al.: Publicly available clinical BERT embeddings. In: NAACL, pp. 72–78 (2019).
33. Peng, Y., Yan, S. and Lu, Z.: Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: 18th BioNLP Workshop and Shared Task, pp. 58–65 (2019).
34. PubMed Central (PMC), https://www.ncbi.nlm.nih.gov/pmc/
35. Transformers library, https://huggingface.co/transformers/
36. word2vec, http://code.google.com/p/word2vec/
37. BMJ Best Practice "COVID-19", https://bestpractice.bmj.com/topics/en-gb/3000168
38. Jurafsky, D. and Martin, J.H.: Speech and language processing (3rd ed. draft). December 2020. https://web.stanford.edu/~jurafsky/slp3/
39. Levy, O., Goldberg, Y.: Linguistic Regularities in Sparse and Explicit Word Representations. In: ACL'14. doi: 10.3115/v1/w14-1618 (2014).
40. Guidotti, R., et al.: A survey of methods for explaining black box models. ACM, doi: 10.1145/3236009, pp.1-42 (2018).
41. COVID-19 specialty guides, https://www.england.nhs.uk/coronavirus/secondary-care/other-resources/specialty-guides/
42. BMJ Best Practice "Management of coexisting conditions in the context of COVID-19", https://bestpractice.bmj.com/topics/en-gb/3000190
43. Viera, A.J., Garrett, J.M.: Understanding interobserver agreement: the kappa statistic. Fam Med. 37(5), pp. 360-3 (2005).
44. HermiT Reasoner, http://www.hermit-reasoner.com/