Ontology Learning with Deep Learning: a case study on Patient Safety using PubMed

M. Arguello Casteleiro[1], M.J. Fernandez-Prieto[2], G. Demetriou[1], N. Maroto[3], W. Read[1], D. Maseda-Fernandez[4], J. Des-Diz[5], G. Nenadic[1], J. Keane[1], and R. Stevens[1]

[1] School of Computer Science, University of Manchester, UK
[2] Salford Languages, University of Salford, UK
[3] Dpto. de Filología Inglesa, Universidad Autónoma de Madrid, Spain
[4] Midcheshire Hospital Foundation Trust, NHS England, UK
[5] Hospital do Salnés, Villagarcía de Arousa, Spain

Abstract. Traditional distributional semantic models (DSMs) like Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) derive representations for words assuming words occurring in similar contexts will have similar representations. Deep Learning has made feasible the derivation of word embeddings (i.e. distributed word representations) from corpora of billions of words applying neural language models like CBOW and Skip-gram. The application of Deep Learning to aid ontology development remains largely unexplored. This study investigates the performance of LSA, LDA, CBOW and Skip-gram for ontology learning tasks. We conducted six experiments; firstly using 300K and later with 14M PubMed titles and abstracts to obtain top-ranked candidate terms related to the *patient safety* domain. Based on the evaluation performed, we conclude that Deep Learning can contribute to ontology engineering from the biomedical literature.

Keywords. Ontology Learning, Deep Learning, OWL-DL, CBOW, Skip-gram

1 Introduction

The World Health Organization (WHO) acknowledges: "*unsafe medication practices and medication errors are a leading cause of injury and health care associated harm around the world*" [1]. In the UK, a report commissioned by the Department of Health states: "*the NHS wastes at least £1bn – and possibly as much as £2.5bn – on preventable errors, many of which are related to improper use of medication*" [2].

In 2009, the WHO proposed an International Classification for Patient Safety (ICPS) with 48 concepts to set forth a common understanding of patient safety literature [3]. A better understanding of the biomedical literature should bring forward more concepts and relations about patient safety. A better understanding of patient safety could affect decisions regarding patients' well-being as well as the expenditure of public funds.

The annotation of concepts and relations from the biomedical literature is key to unlocking the biomedical knowledge contained therein, although the size and rate of

growth of PubMed is a challenge. Recent advances in artificial neural networks (ANNs) make feasible the derivation of words from corpora of billions of words. Hence, the growing interest in Deep Learning, which is an emerging area of artificial neural networks, and the neural language models CBOW (Continuous Bag-of-Words) and Skip-gram of Mikolov et al. [4]. Both CBOW and Skip-gram can efficiently generate word embeddings (i.e. distributed word representations); however, like other methods of distributional semantics, they do not provide the precise formal descriptions that can be found in an ontology.

The manual building of ontologies is often a tedious and cumbersome task [5]. The (semi-)automatic support of ontology development is known as ontology learning [6]. Ontology learning has been depicted as a layer cake [6], where different tasks can be distinguished: a) the acquisition of terms that refer to specific concepts (Named-entity recognition, a.k.a. NER); b) the recognition of synonyms among these terms; c) the identification of taxonomic (is-a) relations; d) the establishment of non-hierarchical relations; and e) the derivation of new knowledge, i.e. knowledge that is not explicitly encoded by the ontology.

This study investigates how CBOW and Skip-gram can be used to aid ontology development for patient safety using PubMed citations (titles and abstracts) as a corpus. The application of Deep Learning to ontology learning tasks, such as concept extraction and relation extraction by Deep Learning, remains largely unexplored.

1.1 Background and Related Work

Distributional semantic models (DSMs) derive representations for words in such a way that words occurring in similar contexts will have similar representations. Therefore, the context needs to be defined. Latent Semantic Analysis (LSA) [7] is a spatially motivated method that generally uses an entire document as a context (i.e. word-document models). Latent Dirichlet Allocation (LDA) [8] is a probabilistic method. Both spatial and probabilistic methods of distributional semantics allow the estimation of the similarity between terms: spatial DSMs compare terms using distance metrics in high-dimensional space [9]; and probabilistic DSMs measure similarity between terms according to the degree to which they share the same topic distributions [9]. LSA and LDA have high computational and storage cost associated with building the model or modifying it due to the huge number of dimensions when a large corpus is modelled [10]. Although neural models are not new in DSMs, in relatively short time, the neural language models CBOW and Skip-gram have gained much popularity to the point of being used for benchmarking word embeddings [11] or as baseline models for performance comparisons [12]. For CBOW and Skip-gram the most popular similarity measure is the cosine of the angle between two vectors of n dimensions.

Much of the work in ontology learning has strong connections with natural language processing and machine learning, and over time, different methods have been applied to learn ontologies and ontology-like structures from text. Indeed, traditional DSMs have been applied already. For example: Colace et al. [13] have used LDA for ontology learning. However, as of today and to the best of our knowledge, CBOW and Skip-gram have not been used for ontology learning. In spite of this, closely re-

lated work can be found: a) Peng et al. [14] use CBOW to generate word embeddings and address automatic MeSH indexing, i.e. multi-class classification; and b) De Vine et al. [15] use Skip-gram to learn word embeddings over sequences of UMLS (Unified Medical Language System) medical concepts instead of over sequences of terms.

2 Materials and Methods

2.1 The Patient-Centric Care (PCC) ontology in OWL-DL

We used the UMLS Metathesaurus 2016AA release from the U.S. National Library of Medicine (NLM) that contains more than three million biomedical related concepts along with synonymous names and their relationships from around 100 source vocabularies, some in multiple languages. Each UMLS concept has a unique identifier (a.k.a. CUI) and is assigned to at least one UMLS Semantic Type [16]. A UMLS Semantic Group contains a set of UMLS Semantic Types [16]. We programmatically created the USTG ontology in OWL-DL [17], which represents formally Semantic Types and Groups as well as the part-whole relations among them. The USTG ontology also contains the UMLS Metathesaurus concept, an OWL class we created that can have as subclass any concept from UMLS. The USTG ontology contains a total of 585 axioms (class count: 151; individual count: 0) and its Description Logic expressivity is ALEI.

To create the Patient-Centric Care (PCC) ontology, we followed the three stages proposed by CommonKADS to construct a knowledge model [18]:

1. *Knowledge Identification* – Besides the USTG ontology, a few information sources were carefully chosen: a) the ontology network about Patient Safety Incident [19] in OWL-DL [17] that is publically available from BioPortal [20] and was supported by WHO under the International Classification for Patient Safety programme; b) the WHO's ICPS [3]; and c) a paper from Mitchell et al. [21] that discusses the structure-process-outcome model of healthcare and acknowledges that outcomes of care should not be limited to what Lohr [22] termed as "The 5Ds" (i.e. death, disease, disability, discomfort, and dissatisfaction), and thus, more positive healthcare outcomes (e.g. improved health status) should also be included. Whenever possible, key concepts from the WHO's ICPS [3] and the ontology network about Patient Safety Incident [19] are mapped to UMLS concepts [16]. A conceptual diagram was drafted.

2. *Knowledge Specification* – We used OWL-DL [17] to formally represent concept names, and concept expressions, along with terminological axioms.

3. *Knowledge Refinement* – we performed some knowledge adjustments and used the FaCT++ reasoner [23] to check logical consistency and concepts' satisfiability. This version of the PCC ontology contains a total of 804 axioms (class count: 181; individual count: 0) and its Description Logic expressivity is ALEHI.

2.2 Ontology Learning with Distributional Semantics

In this study, we adopted lemon (Lexicon Model for Ontologies) [24] and the principle of "semantics by reference" [25]. This principle implies that "the expressivity and the granularity at which the meaning of words can be expressed depend on the meaning distinctions made in the ontology" [25]. Lemon is represented in RDFS [26] and represents lexical information relative to an ontology to make the lexical information shareable on the Semantic Web.

The simplest way to attach a lexical form to an ontological concept is the label property of RDFS (i.e. rdfs:label). In SKOS [27] there are three properties (i.e. skos:prefLabel, skos:altLabel, and skos:hiddenLabel), which can be considered annotation properties (i.e. owl:AnnotationProperty), and only allow limited linguistic information. To include linguistic information more easily, we adopt the following core concepts (OWL classes) and properties (OWL object properties) from the lemon ontology [24]: 1) the OWL classes *Lemon Element*, *Lexicon*, *Lexical entry*, *Lexical Sense*, and *Lexical Topic*; 2) the OWL object properties *entry*, *sense* and its inverse *senseOf*, *reference* and its inverse *isReferenceOf*, and *topic*. We created the LEUSTG ontology that reuses the UTSG ontology and the just mentioned OWL classes and OWL object properties from lemon.

In the LEUSTG ontology, the concept *Lexicon* from lemon represents a vocabulary for a DSM; while the concept Lexical entry from lemon represents a single word (one or more terms) in the vocabulary/lexicon of the DSM. In the LEUSTG ontology, the concept Lexical sense from lemon is superclass of the UMLS Metathesaurus concept from the USTG ontology**.** Hence, any subclass of the OWL class UMLS Metathesaurus concept is a UMLS concept and also an OWL class with one or more UMLS Semantic Types. As UMLS Semantic Types define the senses or meanings of a lexical entry in relation to the given ontology, we follow the "semantics by reference" principle from [25]. In this way these axioms provide a form of class description.

DSMs can facilitate concept extraction and relation extraction (RE) to extend different parts of the ontology. RE has been defined as "*the task of detecting and classifying semantic relations that hold between different entities*" [28]. An overview of the two tasks performed in this study is the following:

Lexical entry and concept extraction – a vocabulary/lexicon from DSMs contains lexical entries that are: concepts, phraseological expressions (typically a combination of concepts), or *spurious* terms (i.e. terms that do not have a true biomedical meaning). UMLS MetaMap [29] – a well-known biomedical NER system by NLM – can indicate which terms from the lexicon are UMLS concepts by assigning them a CUI and also one or more UMLS Semantic Type(s). It should be noted that *patient safety* is a domain not well covered in UMLS where even key concepts such as "*organizational outcome*" do not have a CUI. Hence, some concepts need to be assigned to a UMLS Semantic Type manually.

Extraction of relations – finding association relationships among a large set of terms can be the bases for knowledge discovery. Using the similarity measures (e.g. cosine value for CBOW and Skip-gram) we can quantify empirically how closely related are two terms obtained from the DSM. Once the n top-ranked candidate terms

have been obtained for a query term, it is possible to relate its respective concepts by adding axioms using the ontological constructor skos:related. However, exploiting the knowledge captured in the LEUSTG ontology, this broad relationship is refined. For example, using queries in the SPARQL [30] query language we can easily know: a) term variants of a UMLS concept by querying for candidate terms assigned to the same sense; b) all the candidate terms belonging to the same UMLS Semantic Type or Semantic Group by exploiting the formal relationships between UMLS concepts, UMLS Semantic Types and UMLS Semantic Groups.

Two annotation properties were introduced in the LEUSTG ontology to capture: a) the agreed assessment made by the human raters per candidate term; and b) to what extent the candidate term was recognised by UMLS MetaMap. The LEUSTG ontology contains a total of 691 axioms (class count: 160; individual count: 8) and its Description Logic expressivity is ALEHI.

2.3 Experimental Setup

We downloaded the MEDLINE/PubMed baseline files for 2015 and also the update files up to 8th June 2016. Two biomedical unannotated corpora were obtained: 1) a subset of 301,202 PubMed publications (titles and abstracts) with date of publication from 2000 to 2016 (called here *PubMed SB* for short) obtained by the PubMed Systematic Reviews filter [31]; and 2) 14,056,762 PubMed publications (titles and abstracts) with date of publication from 2000 to 2016 (called here *PubMed 14M* for short). When pre-processing the textual input for CBOW and Skip-gram [4], it is common practice to transform the text into lower-case and to remove all numbers and punctuation marks systematically. This is, however, unsuitable when dealing with protein/gene names, symbols or abbreviations due to the fact that capitalisation and numerals are essential features of their nomenclature. Therefore, we decided to alter the commonly used pre-processing workflow.

Recently Hu et al. [12] experimented with introducing Part-Of-Speech Tagging (POS) information into a neural network similar to CBOW in order to improve the quality of the word embeddings generated. Inspired by the experiments of Hu et al. [12], we pre-processed PubMed citations (titles and abstracts) using two types of approaches: 1) The first approach preserves capitalisation and numbers in the text; and 2) The second approach applies POS tagging and chunking (a.k.a. shallow parsing) to the results of (1). Chunking aims to label segments of a sentence with syntactic constituents, such as noun phrase (NP), and verb phrase (VP).

LDA, LSA, CBOW, and Skip-gram are applied as methods of distributional semantics, where each of them allows the estimation of similarity between terms. We used gensim [32] as the code implementation for LDA and LSA; and word2vec [33] as the code implementation for CBOW and Skip-gram. The terms "*safety*" and "*patient safety*" are the query terms (i.e. the topics). Using similarity metrics (see subsection 1.1) a list of n top-ranked candidate terms can be obtained for each query term. Two domain experts assessed the relevance of the terms in pairs (query term and candidate term) using a Likert-type (categorical) scale taken from [34], which was initially used for patients to capture their level of pain. According to this scale, a candidate

term can be: not at all relevant (marked as 0); a little relevant (marked as 1); quite a bit relevant (marked as 2); and very much relevant (marked as 3). Simple guidelines were given to the domain experts that performed the rating. They consist of: a) the Likert-type (categorical) scale; b) a conceptual diagram that captures the domain of interest; and c) a few examples illustrating pairs of *query term-candidate term* annotated with different scores.

The inter-annotator agreement is calculated with weighted Cohen's Kappa [35], a well-known measure for inter-annotator agreement on classification tasks. Biemann [36] acknowledges: "*for complicated tasks like ontology learning, a comparably low inter-annotator agreement can be expected*".

Brank et al. [37] distinguish three approaches for evaluation in ontology learning depending on the type of ontologies being evaluated and the purpose of the evaluation: 1) *task-based evaluation* using conventional measures in information retrieval [38] like precision, recall, and F-measure; 2) *corpus-based evaluation*; and 3) *criteria-based evaluation*. This study focused on the first two: 1) *task-based evaluation* – a quantitative evaluation was performed in a straight-forward manner applying the well-established lexical precision metric, which measures the number of relevant candidate terms retrieved divided by the total number of candidate terms. In this study, *relevant candidate terms* are terms scored 1 to 3 by both human raters (i.e. domain experts); 2) *corpus-based evaluation* – an evaluation performed at the conceptual level, where the automatically extracted ontology is compared with a Gold Standard ontology that is manually built. In this study, the Gold Standard ontology is the PCC ontology described in subsection 2.1. Three metrics are applied: Lexical Overlap (LO) [39,40] measures the shared concepts between the manual and extracted ontology; Ontological Improvement (OI) [41] accounts for the newly discovered concepts that are absent in the Gold Standard; and Ontological Loss (OL) [41] determines the concepts that exist in the Gold Standard but were not discovered. LO can be interpreted as a "recall" metric. SPARQL [30] queries aided to obtain LO, OI, and OL.

3 Results

Computational resources and execution times – The DSMs are generated using a Supermicro with 256GB RAM and two CPUs Intel Xeon E5-2630 v4 at 2.20GHz. For PubMed dataset SB the execution time goes from less than 1 hour for CBOW (the quickest) to more than 23 hours for LDA (the slowest). For PubMed dataset 14M the execution time goes from less than 1 hour for CBOW to more than 10 hours for Skip-gram. With a MacBook Pro Retina (2.8 GHz Intel Core i7 and 16GB RAM) the mean time for executing each SPARQL query three times was less than 2 seconds.

Distributional Semantics: LSA, LDA, CBOW, and Skip-gram – For LDA and LSA the number of topics was setup to 300, which produced optimal results for similar tasks. For LSA and LDA the top-ranked candidate terms were not always obtained for the query term (i.e. the topic) "*patient safety*". For CBOW and Skip-gram we set up the parameters within the range studied by De Vine et al. [15]. For more technical details refer to the availability note at the end of the paper.

Human Evaluation – A total of 675 pairs of terms (i.e. query term, candidate term) were evaluated. The inter-annotator agreement between the two raters (i.e. domain experts) was 0.62 applying the weighted Cohen's Kappa measure [35].

In figure 1, each column corresponds to a model, and the model's name appears at the top of the column. Each column can have up to three different colours, where each colour depicts the number of candidate terms that were agreed by both raters: dark grey for *very much relevant* (relevant with score equals 3), grey for *relevant* (score 1 or 2), and white for *not relevant* (score 0). Figure 1 on the left hand-side: experiments I, II, III, and IV were performed using the PubMed SB dataset and the maximum number of top candidate terms per model was 20. Figure 1 on the right hand-side: experiments V and VI were performed using the PubMed 14M dataset and the 100 top-ranked candidate terms per model were obtained. In experiments II, IV, and VI noun phrase (NP) and verb phrase (VP) chunking was performed. For experiments I and II the query term (i.e. the topic) was "*safety*"; and for experiments III to VI the query term was "*patient safety*".

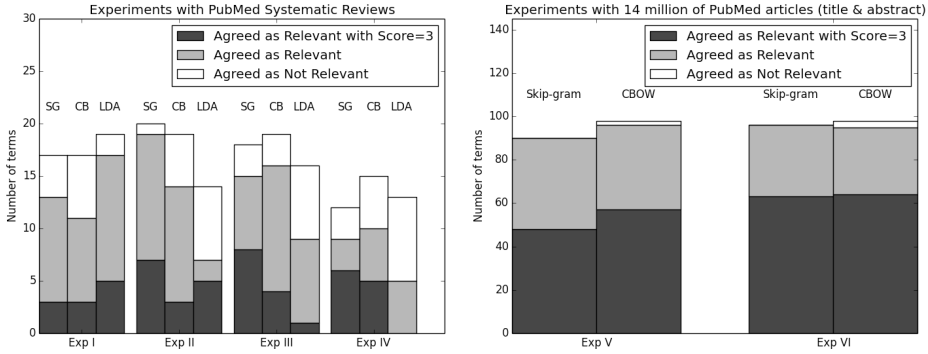

Fig. 1. Agreed assessment by both raters of candidate terms for the different experiments performed. Abbreviations: SG = Skip-gram; CB = CBOW; LD = LDA; Exp = experiment

Table 1 shows the lexical precision calculated as tp/(tp+fp) for each model, where tp stands for the number of true positive agreed by both raters and fp stands for the number of false positive agreed by both raters. For experiments III and IV, there were no top-ranked candidate terms obtained for LSA. After observing the substantial drop in the lexical precision for LDA with the query term "*patient safety*" in experiments III and IV, only the neural language models CBOW and Skip-gram from Deep Learning were considered for experiments V and VI.

Table 1. Lexical precision per model and experiment (abbreviated as Exp.)

Model	Exp. I	Exp. II	Exp. III	Exp. IV	Exp. V	Exp. VI
LSA	64%	45%	-	-	-	-
LDA	**90**%	50%	56%	39%	-	-
CBOW	65%	64%	**84**%	67%	98%	97%
Skip-gram	76%	**95**%	83%	**75**%	**100**%	**98**%

Ontology Learning with Distributional Semantics – the automatically extracted ontology (called here *PubMed Ontology LEarning Ontology* or POLEO for short) refers to the OWL-DL ontology built programmatically out of the top-ranked candidate terms obtained for each model (i.e. LSA, LDA, CBOW, and Skip-gram) in experiments I to VI. The POLE ontology is the result of two tasks: NER and RE (see subsection 2.2) and, it reuses the LEUSTG ontology. The POLE ontology contains a total of 8392 axioms (class count: 689; individual count: 812) and its Description Logic expressivity is ALEHI(D).

Evaluation of Ontology Learning with Deep Learning – Based on the Lexical Precision obtained for experiments I to IV (see Table 1), it is clear that: 1) the neural language models from Deep Learning (i.e. CBOW and Skip-gram) outperformed LDA for "patient safety" as query term (experiment III and IV); and 2) overall Skip-gram seems to have the better Lexical Precision for the experiments conducted.

Another two experiments (V and VI) were set-up using CBOW and Skip-gram only with the PubMed 14M dataset to obtain the 100-top ranked candidate terms. For these two models, we queried the POLE ontology using SPARQL to obtain per model and experiment: a) the total number terms with UMLS concepts as senses agreed as relevant for both raters; 2) the number of terms with UMLS concepts as senses agreed as relevant for both raters that also appear in the PCC ontology; and 3) the number of terms with UMLS concepts as senses agreed as relevant for both raters that are absent in the PCC ontology. With the numbers obtained from the SPARQL queries, we calculated the three ontology learning metrics for *corpus-based evaluation*: LO, OL, and OI (see subsection 2.3) that appear in Table 2.

Table 2. Lexical Overlap (LO), Ontological Loss (OL), and Ontological Improvement (OI)

	Experiment V			**Experiment VI**		
Model	**LO**	**OL**	**OI**	**LO**	**OL**	**OI**
CBOW	**40%**	60%	**57%**	**35%**	65%	**54%**
Skip-gram	**40%**	60%	40%	26%	74%	44%

4 Discussion and Conclusion

From Table 1 and 2, it is difficult to derive a real benefit from the noun phrase (NP) and verb phrase (VP) chunking. There is overall a drop in the performance of the models for experiment II, IV, and VI when they are compared respectively with the results obtained for experiment I, III, and V. Although Skip-gram achieved a significantly better Lexical Precision in experiment II than in experiment I.

In Table 1, Skip-gram obtained a better Lexical Precision than CBOW for most of the experiments. In Table 2, although the Lexical Overlap (LO) is the same for CBOW and Skip-gram in experiment V; CBOW gets a significantly better Ontological Improvement (OI) for both experiments V and VI. Hence, we cannot determine which of the two neural language models from Deep Learning is more suited for ontology learning.

The similarity in the Lexical Overlap (LO) and Ontological Improvement (OI) for experiment V suggest a disconnection between the PubMed corpus and the WHO's ICPS [3].

Deep Learning opens up *unsupervised* learning with big data, as the training of neural language models like CBOW and Skip-gram can be done automatically from a large unannotated corpus and without high computational and storage cost. Hence, a natural long-term venture for Deep Learning is the biomedical literature, which can be seen as a large unannotated corpus with an increasing rate of growth. This paper illustrates how CBOW and Skip-gram can be used to aid ontology learning tasks for *patient safety* using PubMed citations (titles and abstracts) as a corpus. The novelty of this paper is two-fold: 1) ontology learning using Deep Learning remains largely unexplored; and 2) the focus here is on quality of care, and patient safety, where quality care assessment concepts and models are also taken into account.

5 References

1. WHO Global Patient Safety Challenge, http://www.who.int/patientsafety/campaigns/en/
2. The Pharmaceutical Journal, 7834(1), online (2014).
3. Runciman, W., Hibbert, P., Thomson, R., Van Der Schaaf, T., Sherman, H., Lewalle, P.: Towards an International Classification for Patient Safety: key concepts and terms. International journal for quality in health care, 21(1), pp.18-26 (2009).
4. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv pre-print arXiv:1301.3781 (2013).
5. Maedche, A., Staab, S.: Ontology learning. In: Handbook on ontologies, pp. 173-190. Springer, Heidelberg (2004).
6. Buitelaar, P., Cimiano, P. and Magnini, B.: Ontology learning from text: An overview. Ontology learning from text: Methods, evaluation and applications, 123, pp.3-12 (2005).
7. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2), p.211 (1997).
8. Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022 (2003).
9. Cohen, T., Widdows, D.: Empirical distributional semantics: methods and biomedical applications. Journal of biomedical informatics, 42(2), pp.390-405 (2009).
10. Jonnalagadda, S., Leaman, R., Cohen, T. and Gonzalez, G.: A distributional semantics approach to simultaneous recognition of multiple classes of named entities. In: CICLing, pp. 224-235. Springer, Heidelberg (2010).
11. Neelakantan, A., Shankar, J., Passos, A., McCallum, A.: Efficient non-parametric estimation of multiple embeddings per word in vector space. In: EMNLP, pp. 1059-1069 (2014).
12. Hu, B., Tang, B., Chen, Q., Kang, L.: A novel word embedding learning model using the dissociation between nouns and verbs. Neurocomputing, 171, pp.1108-1117 (2016).
13. Colace, F., De Santo, M., Greco, L., Amato, F., Moscato, V., Picariello, A: Terminological ontology learning and population using latent Dirichlet allocation. Journal of Visual Languages and Computing, 25(6), pp. 818-826 (2014).
14. Peng, S., You, R., Wang, H., Zhai, C., Mamitsuka, H. and Zhu, S.: DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. Bioinformatics, 32(12), pp.i70-i79 (2016).

15. De Vine, L., Zuccon, G., Koopman, B., Sitbon, L., Bruza, P.: Medical semantic similarity with a neural language model. In: ACM CIKM, pp. 1819-1822 (2014).
16. UMLS Semantic Network, https://semanticnetwork.nlm.nih.gov
17. OWL 2, https://www.w3.org/TR/2009/REC-owl2-overview-20091027/
18. Schreiber, G.: Knowledge engineering and management: the CommonKADS methodology. MIT press (2000).
19. Ontology network ICPS, https://bioportal.bioontology.org/ontolo-gies/ICPS
20. BioPortal, http://bioportal.bioontology.org
21. Mitchell, P.H., Ferketich, S., Jennings, B.M.: Quality health outcomes model. Image: The Journal of Nursing Scholarship, 30(1), pp.43-46 (1998).
22. Lohr, K.N.: Outcome measurement: concepts and questions. Inquiry, pp.37-50 (1988).
23. FaCT++, http://owl.man.ac.uk/factplusplus/
24. lemon, http://lemon-model.net
25. Cimiano, P., McCrae, J., Buitelaar, P., Montiel-Ponsoda, E.: On the role of senses in the ontology-lexicon. In: New trends of research in ontologies and Lexical resources, pp. 43-62. Springer, Heidelberg (2013).
26. RDFS, https://www.w3.org/TR/2000/CR-rdf-schema-20000327/
27. SKOS, https://www.w3.org/TR/skos-reference/
28. Petrova, A., Ma, Y., Tsatsaronis, G., Kissa, M., Distel, F., Baader, F. and Schroeder, M., 2015. Formalizing biomedical concepts from textual definitions. Journal of biomedical semantics, 6(1).
29. MetaMap, https://metamap.nlm.nih.gov
30. SPARQL 1.1, https://www.w3.org/TR/sparql11-query/
31. PubMed SB, https://www.nlm.nih.gov/bsd/pubmed_subsets/sys reviews_strategy.html
32. gensim, https://radimrehurek.com/gensim/
33. word2vec, http://code.google.com/p/word2vec/. Accessed 12 Nov 2014.
34. Aaronson, N.K.: Quality of life assessment in clinical trials: methodologic issues. Controlled Clinical Trials, 10(4), pp.195-208 (1989).
35. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin, 70(4), p.213 (1968).
36. Biemann, C.: Ontology learning from text: A survey of methods. In: LDV forum, 20(2), pp. 75-93 (2005).
37. Brank, J., Grobelnik, M., Mladenic, D.: A survey of ontology evaluation techniques. In: SiKDD, pp. 166-170 (2005).
38. Wong, W., Liu, W., Bennamoun, M.: Ontology learning from text: A look back and into the future. ACM CSUR, 44(4), p.20 (2012).
39. Maedche, A., Staab, S.: Measuring similarity between ontologies. In: EKAW, pp. 251-263. Springer, Heidelberg (2002).
40. Cimiano, P., Staab, S., Tane, J.: Automatic acquisition of taxonomies from text: FCA meets NLP. In: ECML/PKDD, pp. 10-17 (2003).
41. Sabou, M., Wroe, C., Goble, C., Mishne, G.: Learning domain ontologies for web service descriptions: an experiment in bioinformatics. In: ACM WWW, pp. 190-198 (2005).

Acknowledgement – To Prof Iain Buchan, Chris Wroe, D. Tsarkov and Stephen Walker for useful discussions; and to Timothy Furmston for helping with the software and e-infrastructure.

Funding – This work was supported by a grant from the European Union Seventh Framework Programme (FP7/2007-2013) for sysVASC project under grant agreement number 603288.

Availability – For technical details as well as for the hyperlink to download the ontologies mentioned in the paper, please refer to http://pole-dl.cs.manchester.ac.uk