Medical Concept Embeddings via Labeled Background Corpora

Eneldo Loza Mencía[1], **Gerard de Melo**[2], **Jinseok Nam**[3]
[1] Knowledge Engineering Group, Technische Universität Darmstadt, Germany
[2] IIIS, Tsinghua University, Beijing, China
[3] Knowledge Discovery in Scientific Literature, Technische Universität Darmstadt, Germany
eneldo@ke.tu-darmstadt.de, gdm@demelo.org, nam@kdsl.informatik.tu-darmstadt.de

Abstract

In recent years, we have seen an increasing amount of interest in low-dimensional vector representations of words. Among other things, these facilitate computing word similarity and relatedness scores. The most well-known example of algorithms to produce representations of this sort are the word2vec approaches. In this paper, we investigate a new model to induce such vector spaces for medical concepts, based on a joint objective that exploits not only word co-occurrences but also manually labeled documents, as available from sources such as PubMed. Our extensive experimental analysis shows that our embeddings lead to significantly higher correlations with human similarity and relatedness assessments than previous work. Due to the simplicity and versatility of vector representations, these findings suggest that our resource can easily be used as a drop-in replacement to improve any systems relying on medical concept similarity measures.

Keywords: medical concepts, semantic similarity, MeSH, embeddings

1 Introduction

For many decades, researchers have debated the curious nature of human language. On the one hand, language appears to naturally involve discrete symbolic units. On the other hand, these symbolic units are clearly not independent. There is a clear relationship, for instance, between the two words *neuron* and *neural*. The same applies to multi-word expressions, as are common in the medical domain, for instance, between the *central nervous system* and the *peripheral nervous system*.

In the past, many models ignored such semantic relationships, or invoked custom techniques such as query expansion, to cope with them. Recently, another alternative has proven useful and acquired significant popularity: Discrete symbols can be mapped to a low-dimensional vector space, in which distances or angles between vectors reflect similarities between the symbols. Many of the approaches for this rely on neural network techniques. However, existing approaches such as the word2vec Skip-Gram with Negative Sampling model (Mikolov et al., 2013) neglect valuable additional information that may be available in the used document collections, e.g. human annotations or hierarchical information.

In this paper, we investigate a new approach called All-in-text for producing embeddings for medical concepts (Nam et al., 2016). All-in-text was originally proposed for keyword indexing and multi-label classification. In this paper, we adapt it to the task of producing embeddings for medical concepts. The learning method is inspired by the Paragraph Vector technique (Le and Mikolov, 2014) of learning representations of words and word sequences (documents), which was originally used for text classification and sentiment analysis. However, All-in-text is a versatile framework that can incorporate more complex connectivity information originating from the associations between documents and target concepts in a background corpus. Fig. 1 shows some example documents from the used corpus, associated with the concepts *Malnutrition* and *Overnutrition*.

Our experimental evaluation show that this approach yields state-of-the-art results on two medical semantic similarity and relatedness datasets. Moreover, our vector representations can also be used in more flexible ways than standard measures, as these representations encode additional kinds of semantic information that can directly be exploited as features by neural networks.

2 Related Work

In the past, coping with the semantic similarity of words often meant relying on custom lexical resources. In the simplest case, this could be a simple list of synonyms or aliases. Lexical networks such as WordNet (Fellbaum, 1998) led to extensive research on more sophisticated methods that exploited graph connectivity and gloss comparisons.

In the medical domain, similar lexical resources and ontologies have been created. Major examples are the Medical Subject Headings (MeSH)[1], the Unified Medical Language System (UMLS), and the SNOMED clinical terms (SNOMED CT). Similarity measures such as the one proposed by Wu and Palmer (1994) or Nguyen and Al-Mubaid[2] consider the depth in the hierarchical structures of the used ontologies or path lengths between two concepts in order to compute a similarity metric. Resnik (1995) as well as Jiang and Conrath (1997) propose to additionally exploit the occurrence probabilities of the concepts, as computed on large corpora.

Less ontology-dependent measures such as the Lesk measure and the Vector approach from Liu et al. (2012) rely on textual descriptions of the concepts and context expansions to compute the relatedness between terms. More details on such measures, especially on the relatedness measures, are given by Liu et al. (2012).

Another line of work proposed more data-driven methods without the need for a knowledge base. The most well-

[1] https://www.nlm.nih.gov/mesh/

[2] http://atlas.ahc.umn.edu/umls_similarity/similarity_measures.html

Title: Spotlight on Global Malnutrition: A Continuing Challenge in the 21st Century.
Abstract: Malnutrition as undernutrition, overnutrition, or an imbalance of specific nutrients, can be found in all countries and in both community and hospital settings around the world. The prevalence of malnutrition is unacceptably high ...
MeSH terms: Acute Disease, Chronic Disease, Food Habits, Global Health, Humans, Malnutrition, Nutritional, Support, Overnutrition, Risk Factors, Socioeconomic Factors

Title: Fetal and early-postnatal developmental patterns of obese-genotype piglets exposed to prenatal programming by maternal over- and undernutrition.
Abstract: The present study evaluated the effect of nutritional imbalances during pregnancy, either by excess or deficiency, on fertility and conceptus development in obese-genotype swine (Iberian pig). Twenty-five multiparous sows were ...
MeSH terms: Animals, Newborn Animals, Body Weight, Fetal Development, Genotype, Malnutrition, Obesity, Overnutrition, Pregnancy, Prenatal Exposure Delayed Effects, Swine

Title: Predictors of maternal and child double burden of malnutrition in rural Indonesia and Bangladesh
Abstract: BACKGROUND: Many developing countries now face the double burden of malnutrition, defined as the coexistence of a stunted child and overweight mother within the same household. OBJECTIVE: This study sought to ...
MeSH terms: Adult, Body Mass Index, Preschool Child, Cost of Illness, Cross-Sectional Studies, Developing Countries, Family Characteristics, Humans, Indonesia, Infant, Logistic Models, Malnutrition, Mothers, Overnutrition, Population Surveillance, Prevalence Risk Factors, Rural Health, Urban Health

Figure 1: Example entries from the BioASQ dataset of PubMed abstracts and associated MeSH terms. The combination *Malnutrition – Overnutrition* appears 43 times in the dataset.

known family of such methods, aiming at overcoming the discreteness of symbols, are Latent Semantic Analysis or LSA (Deerwester et al., 1990), which applies singular value decomposition for dimensionality reduction of the term-document matrix, and its Bayesian probabilistic descendant Latent Dirichlet Allocation or LDA (Blei et al., 2003). These methods have been very influential in natural language processing and information retrieval. Still, many of them suffer from limited scalability and normally need to be re-applied to new document collections. Distributional semantic methods rely on term co-occurrence matrices rather than term-document matrices (Schütze, 1993), delivering quite meaningful results. However, experimental results suggest that newer neural network-based models produce better word representations than both LSA and traditional distributional semantic methods (Pennington et al., 2014).
In recent years, low-dimensional embeddings have been proposed as a particularly simple way to feed such knowledge of similarities into machine learning algorithms (Collobert et al., 2011; Turian et al., 2010). The fast algorithms by Mikolov et al. (2013) and their freely available word2vec implementation[3] as well as the publicly available pretrained data has made such word embeddings very convenient to use. In recent years, numerous extensions have been proposed, e.g. better exploiting information extraction pattern occurrences (Chen and de Melo, 2015) or multilingual structured data (de Melo, 2015). All-in-text is based on the scalable Paragraph Vector algorithm (Le and Mikolov, 2014), which extends the word2vec ideas to jointly create representations of words and word sequences such as sentences, paragraphs, or entire documents.

3 All-in-text for Medical Concepts

Our goal is to learn representations for medical concepts. In particular, we would like to project a given concept y to a k-dimensional vector $\mathbf{v_y} \in \mathbb{R}^k$. The size of our representation space $\mathbb{R}^k$ will typically be in the order of hundreds, and thus much lower than in traditional term-vector spaces, where the dimensionality corresponds to the size of the vocabulary.

Many neural approaches to learning such vector representations are based on the idea that the vectors for a series of words, taken as input, should enable the *prediction* of related words such as those co-occurring in some context window. Using gradient-based optimization, one can keep altering the vectors so as to facilitate such predictions. The Paragraph Vector algorithm by Le and Mikolov (2014) extended this idea to create vector representations of entire documents (or paragraphs). The vector representation for a given document is included as part of the input that is used to predict words.

The All-in-text approach draws on this framework to jointly derive vector representations of documents as well as vector representations of class labels associated with such documents (Nam et al., 2016). The approach relies on a corpus of texts, in which each document has a (manually created) set of labels, in our case with medical concepts. The objective of the learning algorithm is to jointly learn vector representations for documents and labels, exploiting the label assignments for a given document.

The original purpose of the approach was automatic keyword indexing or multi-label classification, i.e., the model can be used to infer a list of relevant class labels for an unseen test document by computing a distance score between the document representation and all its label representations. A good model will ensure a high compatibility between documents and their associated labels. Based on the assumption that related or similar class labels tend to co-occur in documents more often than unrelated labels , we expect that vector representations for similar labels will tend to be closer to each other due to the associations via documents for which such labels co-occur. However, relationships of this sort could conceivably also be discoverable simply by counting co-occurrences of class labels. Indeed, our experimental results in Section 5 show that this turns out to already be a

[3] https://code.google.com/p/word2vec/

strong baseline. However, this idea only works for labels that have been observed together. By jointly embedding documents and labels, we are able to discern associations between labels that are not directly observed together, but can be detected indirectly via associated documents, for example due to similar labels or similar content.

In our case, we consider documents labeled with medical concepts and hence both medical documents and medical concept labels are jointly embedded. We capture similarities between concepts via document representations, exploiting both documents showing direct co-occurrences of concept labels as well as indirect connections, e.g. via documents with similar terms within them. In the following, we describe the All-in-text algorithm (also called *AiTextML*, or All-in-text joint embeddings for multi-label classification), focusing on the parts that were used in our case. We refer the reader to Nam et al. (2016) for more details, e.g. on the additional ability of the approach to learn from textual class descriptions.

3.1 Word and Document Embeddings

Assume that we are given a vocabulary of V words $\mathcal{W} = \{1, 2, \ldots, V\}$, a set of concepts $\mathcal{C} = \{1, 2, \ldots, L\}$, and a set of N training examples $\mathcal{D} = \{(\mathcal{T}^{(i)}, \mathcal{Y}^{(i)})_{i=1}^{N}\}$ where $\mathcal{T}^{(i)} = \{w_1^{(i)}, w_2^{(i)}, \cdots, w_{|\mathcal{T}^{(i)}|}^{(i)}\}$ denotes a sequence of $|\mathcal{T}^{(i)}|$ words $w_j^{(i)} \in \mathcal{W}$, and $\mathcal{Y}^{(i)} = \{y_1^{(i)}, y_2^{(i)}, \cdots, y_{|\mathcal{Y}^{(i)}|}^{(i)}\}$ is the set of relevant labels $y_j^{(i)} \in \mathcal{C}$ for the i-th training example. AiTextML learns vector representations $\mathbf{U} = \{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_V\} \in \mathbb{R}^{k \times V}$ for the words in $\mathcal{W}$, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \in \mathbb{R}^{k \times N}$ for training documents $\mathcal{T}^{(i)}$, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_L\} \in \mathbb{R}^{k \times L}$ for labels $\mathbf{y}_i$, and $\mathbf{U}' = \{\mathbf{u'}_1, \mathbf{u'}_2, \cdots, \mathbf{u'}_V\} \in \mathbb{R}^{c \cdot k \times V}$ for word contexts, where d is the desired embedding dimensionality and c is the size of the context window.

We use the objective function of the Paragraph Vector algorithm in order to learn the connection between document and word representations, namely to maximize the probability $p(w_t \mid \mathbf{w}_{-t}, \mathbf{x})$ of predicting a word w_t at a certain position t in a document $\mathcal{T}$, given its surrounding words $\mathbf{w}_{-t}$ and the representation of the document (Le and Mikolov, 2014). More specifically, this probability is given by

$$p(w_t \mid \mathbf{w}_{-t}, \mathbf{x}) = \frac{\exp(\mathbf{u'}_{w_t}^T \hat{\mathbf{u}}_{w_t})}{\sum_{v=1}^{V} \exp(\mathbf{u'}_v^T \hat{\mathbf{u}}_{w_t})} \quad (1)$$

where $\mathbf{u}'_{w_t}$ is the $c \cdot k$-dimensional vector for a central (output) word w_t, and the context $\hat{\mathbf{u}}_{w_t}$ of w_t is given by the concatenation of context word embeddings $\mathbf{w}_{-t} = \{w_{t-(c-1)/2}, \ldots, w_{t-1}, w_{t+1}, \cdots, w_{t+(c-1)/2}\}$ as well as of the document embedding $\mathbf{x}$, as defined by

$$\hat{\mathbf{u}}_{w_t} = \left[\mathbf{x}, \mathbf{u}_{w_{t-(c-1)/2}}, \cdots, \mathbf{u}_{w_{t+(c-1)/2}}\right] \in \mathbb{R}^{c \cdot k}. \quad (2)$$

As computing the denominator becomes intractable for large vocabularies, we use negative sampling for efficiently approximating the softmax formulation (Mikolov et al., 2013). Hence, we approximate the logarithm $\log p\left(w_t \mid \mathbf{w}_{-t}, \mathbf{x}\right)$ of the probability by sampling only κ words out of $\mathcal{W}$, resulting in

$$\log \sigma(\mathbf{u'}_{w_t}^T \hat{\mathbf{u}}_{w_t}) + \sum_{j=1}^{\kappa} \mathbb{E}_{P_n(w)} \left[\log \sigma(\mathbf{u'}_{w_j}^T \hat{\mathbf{u}}_{w_t})\right] \quad (3)$$

where $\sigma(x)$ is the sigmoid function, and $P_{n(w)}$ is the frequency distribution of the words in the corpus raised to the power of $3/4$.

3.2 Label Embeddings

Until now, modeling the relationship between documents and their concept labels is disregarded. However, since our goal is to maintain the relationship structure between documents and their labels in the background corpus, for a given document $\mathcal{T}^{(i)}$ and its associated representation $\mathbf{x_i}$, we learn to place the embeddings of associated concept labels $\mathcal{Y}^{(i)}$ closer to $\mathbf{x}$ than for the negative labels $\bar{\mathcal{Y}}^{(i)} = \mathcal{C} \backslash \mathcal{Y}^{(i)}$. More formally, our new objective is to minimize the ranking loss

$$\sum_{y^+ \in \mathcal{Y}^{(i)}} \sum_{y^- \in \bar{\mathcal{Y}}^{(i)}} \mathbb{I}\left[f\left(\mathbf{x_i}, \mathbf{y}_{y^+}\right) \leq f\left(\mathbf{x_i}, \mathbf{y}_{y^-}\right)\right] \quad (4)$$

where $\mathbb{I}[\cdot]$ takes 1 if its argument is true otherwise 0, and $f(\mathbf{x}, \mathbf{y})$ denotes the similarity between the respective representations of document $\mathcal{T}$ and label y. It is computed by

$$f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{W} \mathbf{y}. \quad (5)$$

where $\mathbf{W}$ denotes a bilinear mapping between the label and the document space, which is also learned by the algorithm. For efficiency reasons, we use the weighted approximate rank pairwise (WARP) variant of Eq. 4 (Weston et al., 2011). Roughly speaking, similarly to negative sampling, WARP samples negatives labels $y^- \in \mathcal{Y}^{(i)}$ until it finds a negative label which was correctly ranked above the corresponding positive label. Eq. 4 is then estimated by computed a weighted sum over the differences in the distances $m - f(\mathbf{x}, \mathbf{y}_{y^+}) + f(\mathbf{x}, \mathbf{y}_{y^-})$ (the extent of error of ranking the negative label above the positive one) where m is a user-defined margin (in our case $m = 0.1$).

Stochastic gradient descent with a fixed learning rate is used to train the parameters. The overall objective function is given by adding up the sum over the negative logs (Eq. 3) for each word in the corpus and the approximation of the rank loss (Eq. 4) for each positive label in each of the documents. Both terms can be weighted by two parameters α and β, respectively. Technically, in each epoch, AiTextML iterates over all documents, where it first updates the representations of the sampled positive and negative labels, and then updates the word representations.

3.3 Distance Computation

The part of the formulation of the problem for finding word embeddings is very similar to the continuous bag of words model (Mikolov et al., 2013), with the main difference that a context document is added. Therefore, the resulting word vector representations $\mathbf{u}$ can naturally be used for distance computations by computing the cosine similarity or Euclidean distance between a pair of embeddings.

In the same manner, we can compute distances between the vector representations of concept labels. Based on Eq. 5,

we can also relate document representations to concept label embeddings. Eq. 4 ensures that document embeddings are compatible with their corresponding label embeddings, as well as the other way around, while Eq. 1 ensures that documents with similar words and word sequences are close.

4 Experimental Setup

Table 1: Mismatches of the approaches on the respective datasets: Number of pairs for which no similarity measure could be computed.

Dataset / Approaches	UMNSRS -rel	UMNSRS -sim	MayoSRS
Number of pairs	587	566	101
Nguyen & Al-Mubaid	268	257	50
Path	267	256	50
Wu & Palmer	268	257	50
Lin	267	256	50
Jiang & Conrath	268	257	50
Resnik	267	256	50
Lesk	41	33	14
Vector	41	33	14
PubMed	58	49	65
PMC	96	83	65
Pubmed & PMC	52	43	65
Wkpd & PubMed & PMC	50	42	65
Number of co-occurrences	333	312	51
AiTextML (label)	207	194	38
AiTextML (both)	59	45	27
AiTextML (word)	122	102	68

4.1 Evaluation Datasets

For evaluation, we mainly relied on the Medical Residents Similarity and Relatedness Set datasets (Pakhomov et al., 2010). These two datasets (UMNSRS-sim and UMNSRS-res) consist of concept name pairs and similarity or relatedness assessments, respectively, made by 8 medical residents. In addition, we evaluated our approach on the Medical Coders Set (MayoSRS) of 101 medical concept pairs rated for semantic relatedness by medical coders (Pedersen et al., 2007).

As our evaluation measure, we use the Spearman rank correlation between the human-provided scores and the scores computed by the respective algorithms. This metric measures the non-parametric statistical dependence between two ranked variables. It is independent of the actual absolute similarity scores obtained and is hence commonly used to compare different systems that may compute similarities based on different principles. The correlation coefficient between two rankings is given by

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} r_A(i) - r_B(i)}{n(n^2 - 1)} \qquad (6)$$

where $r_A(i)$ are the ranks of the scores $s_A(i)$ of a method A for concept pairs $i = 1 \ldots n$, and similarly for method B. ρ lies between 1 and 1, where 1 would be a complete positive correlation, 0 means there is no correlation, and 1 would be a complete anti-correlation. We use the standard procedure for handling ties correctly, i.e. tied values are assigned the average of all ranks of items sharing the same value in the ranked list, sorted in ascending order of the values.

Table 2: Evaluation results in terms of Spearman rank correlation on the UMNSRS relatedness dataset for the path, context-expansion, continuous vector representations, and labeled background corpora based approaches (first to fourth blocks, respectively).

Dataset subset and size / Approaches	Smallest 171	Small 241	Middle 309	Largest 440
Nguyen & Al-Mubaid	.3076	.2797	–	–
Path	.3157	.2861	–	–
Wu & Palmer	.3109	.2793	–	–
Lin	.2867	.2868	–	–
Jiang & Conrath	.2989	.2772	–	–
Resnik	.2548	.2999	–	–
Lesk	.3557	.3654	.3023	.3135
Vector	.3859	.4827	.4526	.4650
PubMed	.3384	.3688	.3784	.3757
PMC	.2039	.2288	–	–
Pubmed & PMC	.3289	.3482	.3629	.3563
Wkpd & PubMed & PMC	.3227	.3523	.3496	.3635
Co-occurrences	.5386	–	–	–
AiTextML (label)	.4659	.5500	.5297	–
AiTextML (both)	–	–	–	.5397
AiTextML (word)	.2875	.3194	.2998	.3798

Table 3: Evaluation results on the UMNSRS similarity dataset.

Dataset subset and size / Approaches	Smallest 171	Small 237	Middle 307	Largest 440
Nguyen & Al-Mubaid	.3456	.2689	–	–
Path	.3459	.2685	–	–
Wu & Palmer	.3543	.2753	–	–
Lin	.3361	.2941	–	–
Jiang & Conrath	.3705	.2907	–	–
Resnik	.2845	.3018	–	–
Lesk	.4154	.4140	.3971	.4078
Vector	.4976	.5165	.5008	.5113
PubMed	.4483	.4298	.4426	.4660
PMC	.3409	.3054	–	–
Pubmed & PMC	.4354	.4034	.3985	.4300
Wkpd & PubMed & PMC	.4243	.3912	.3916	.4366
Co-occurrences	.5210	–	–	–
AiTextML (label)	.5910	.5960	.5626	–
AiTextML (both)	–	–	–	.5632
AiTextML (word)	.3890	.3476	.3356	.4183

The terms in the UMNSRS datasets were selected from the UMLS ontology and do not necessarily appear in the MeSH hierarchy used by our method and the baselines. In addition, the corpus-based approaches only cover a part of the concepts due to limitations on the frequency of concepts appearing in the dataset. There are also a few isolated cases of multi-word expressions not identified by the word2vec phrase recognition implementation. Table 1 provides an overview of the concept pairs that could not be processed by the different approaches. For a fair comparison, we focused on different intersections covered by subsets of the methods. Hence, the scores are not directly comparable between the different subsets of concept pairs.

Table 4: Evaluation results on the MayoSRS dataset.

Dataset subset and size /	Smallest	Small	Middle	Largest
Approaches	17	19	21	28
Nguyen & Al-Mubaid	.1395	.1677	–	–
Path	.1044	.1396	–	–
Wu & Palmer	.1830	.2262	–	–
Lin	.2042	.2690	–	–
Jiang & Conrath	.2765	.3411	–	–
Resnik	-.0209	.0432	–	–
Lesk	.6568	.7125	.5660	.4410
Vector	.5092	.5785	.5086	.4948
PubMed	.4846	.5370	.5295	.4827
PMC	.3087	.4171	–	–
Pubmed & PMC	.4022	.4727	.4819	.4525
Wkpd & PubMed & PMC	.3296	.4189	.4278	.4319
Co-occurrences	.3430	–	–	–
AiTextML (label)	.7737	.7910	.7147	–
AiTextML (both)	–	–	–	.6376
AiTextML (word)	.3813	.3016	.3143	.3427

4.2 Training Corpora

We trained the embeddings on 11.6 million documents from the BioASQ Task 3a dataset[4], a large collection of scientific publications on biomedical research extracted until the year 2015 from the PubMed platform.[5] The documents were labeled by the PubMed curators with 26,103 medical concepts from the MeSH ontology, on average 11.2 per document. Only the title and abstracts of the publications were available and were hence used.

For the remaining parameters, we used the default settings of the freely available implementation of AiTextML[6], i.e., a vector size of 100, window size of 5, learning rate of 0.025, 5 negative samples, and 20 epochs. As outlined in Section 3, in our experiments we did not consider learning from label descriptions. Instead, we used only the label to document associations (with a weight of $\frac{1}{3}$) and the interactions between documents and words (with a weight of $\frac{2}{3}$).

We obtained vector representations for 26,103 of the MeSH terms[7] appearing in the documents of the collection. The approach also obtained 513,196 embeddings for the words appearing at least 20 times in the corpus.

The version of the model which only uses label embeddings is referred to as *AiTextML (label)*, whereas *AiTextML (word)* uses only word embeddings for scoring concept pairs. A third variant called *AiTextML (both)* switches to word embeddings only if no corresponding label embedding could be found for one of the concepts.

4.3 Baselines

We compared our results against several state-of-the-art measures as implemented in the well-known UMLS::Similarity package (version 1.41) (McInnes et al., 2009). These measures were briefly introduced in Section 2. The UMLS graph and the included concept descriptions were used for the path and context expansion based measures.

Recently, Pyysalo et al. (2013) computed word vector representation based on the popular word2vec skip-gram with negative sampling approach on large corpora from the medical domain. One of the variants was trained similarly to our approach on almost 23 million article titles and abstracts (*PubMed*). A second variant was computed on nearly 700,000 full article texts from PubMed Central Open Access (*PMC*). Two additional variants aggregated *PubMed & PMC*, on the one hand, and additionally the English Wikipedia on the other hand (*Wkpd & PubMed & PMC*). The authors created 200 dimensional embeddings using the original word2vec implementation with a window size of 5, hierarchical softmax, and a frequent word subsampling threshold of 10^{-3}.[8]

Contrary to our proposed approach, these models make no explicit usage of the label information included with the PubMed data. In order to provide a baseline for using this additional information, we report the results for ranking concept pairs according to their number of co-occurrences in the BioASQ corpus.

5 Experimental Results

We compared the methods on different subsets of the original sets of pairs. The first subset (*Smallest*) was obtained by using the concept pairs that were covered by all approaches. The second one (*Small*) was obtained in the same way, but omitting the approach that scores pairs based on the number of co-occurrences found, as it had a very low coverage (cf. Table 1). The *Middle*-sized subset skips the path-based approaches as well as the *PMC* dataset, whereas the largest subset (*Largest*) was obtained by using the combination of AiTextML label and word embeddings instead of *AiTextML (label)*.

The results are given in Tables 2, 3 and 4 for the corresponding concept pair datasets. Comparing the Spearman rank correlation scores for different measures, the first observation is that graph-based methods (first block) are generally dominated by all other methods (except PMC), regardless of the specific task or dataset. Among the context word co-occurrence approaches (skip-gram models in the third block and also AiTextML (word)), we observe that the correlation depends on the corpus size. PMC uses the smallest corpus, followed by AiTextML (word) and the larger corpora including all PubMed abstracts. However, adding Wikipedia in addition to the corpora from the medical domain generally lowers the quality of the produced scores, indicating that out-of-domain data may not be useful.

Interestingly, the continuous word vector representation models are outperformed by the Vector approach by Liu et al. (2012), which relies on simple discrete vectors for glosses. We conjecture that although such traditional discrete vectors are known to be inferior, the gloss descriptions they are computed from provide valuable explicit semantic information. This indicates that AiTextML could presumably yield even better results by additionally exploiting such glosses, which we did not make use of here in order to keep our method more general.

[4] http://www.bioasq.org/participate/data

[5] http://www.ncbi.nlm.nih.gov/pubmed

[6] https://github.com/JinseokNam/AiTextML

[7] We used the 2015 version from http://www.nlm.nih.gov/mesh/filelist.html

[8] Freely available at http://bio.nlplab.org/.

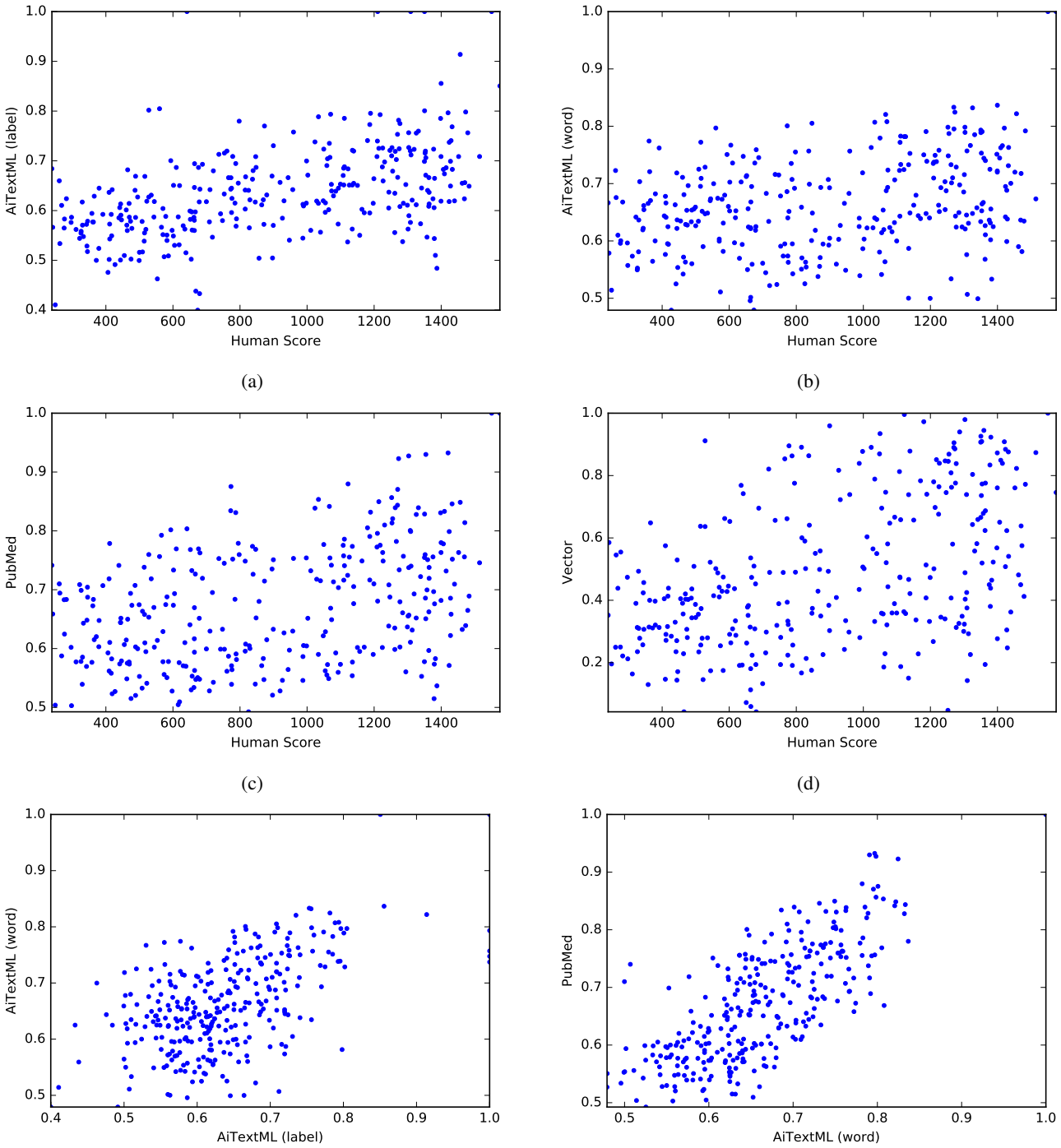

Figure 2: Correlations of scores for UMNSRS-rel. The respective scores are referred to on the axis. Axis are cropped by minimum and maximum values.

Exploiting the additional relationships between concepts and documents with our proposed method AiTextML (label) provides a substantial additional improvement compared to the embedding approaches relying only on word context information. AiTextML (label) and (both) improve the correlation with human annotators for all datasets, also with respect to the strong gloss vector baseline.

Regarding the difference between the assessment of relatedness and similarity, our approach seems to be better suited for evaluating the similarity of concepts. Note that similarity and relatedness are two different concepts, oftentimes with conflicting ground truths, as can be seen in Table 5. Hence, it is not always possible to optimize both simultaneously. In fact for UMNSRS-rel, just using the number of co-occurrences as a metric seems to be the best option in order to evaluate the relatedness. However, the coverage of such a counting-based approach is very low. In contrast, by not only considering co-occurrences, our proposed method can capture relatedness even if two concepts were not observed directly together.

Table 5: Inspection of example medical concept pairs. *Similarity* and *Relatedness* show the respective ranks of the listed concept pairs according to the human evaluation. Ranks start at 0 and there were 167 remaining elements.

Left term	Right term	Relatedness	Similarity	AiTextML (label)	AiTextML (word)	Co-Occurrences
Ethanol	Alcohol	1	7	42	74	21
Medrol	Prednisolone	75	1	22	31	11
Nausea	Vomiting	6	32	1	5	2
Polydipsia	Polyuria	17	22	2	1	102.5
Hypothyroidism	Synthroid	4	39	4	142	1
Angina	Dyspnea	27.5	110	88	36	36
Xanax	Ativan	97	6	9	8	50
Hernias	Earache	165	160	53	78	147
Ataxia	Ethanol	21	78	163	113	41
Overnutrition	Malnutrition	87	165	7	7	66
Cirrhosis	Hematemesis	96	31	147	103	147
Anosmia	Constipation	154	152	122	25	136.5
Pallor	Iron	22	57	159	160	147
Starvation	Anorexia	11	17	77	132	96
Syphilis	Gonorrhea	9	24	16	40	18
Bronchitis	Pneumonia	88	34.5	26	6	14
Carboplatin	Cisplatin	46	5	18	10	4

Fig. 2 shows correlations between different scores on the UMNSRS relatedness dataset. In the ideal case of a perfect correlation, we would observe strictly monotonically increasing curves. Our analysis shows that AiTextML (label) shows significantly better correlation with human assessments (less dispersion) than alternative methods. The AiTextML (label) and (word) variants clearly do not learn the same underlying concept of distances (cf. Fig. 2c). In fact, the word variant is highly correlated with the regular word2vec skip-gram approach on the PubMed dataset (cf. Fig. 2f).

Exploiting the connections between documents and concept labels seems to be less beneficial for the relatedness task, but the increase compared to previous approaches is still very pronounced. A line of future research could be to additionally exploit textual concept descriptions, which are often available in medical ontologies, in order to further improve our results on the relatedness task. Our method can naturally be extended for this purpose, since it provides the means for embedding such descriptions into the same joint vector space in our setting.

5.1 Detailed Analysis

Table 5 lists some examples of medical concept pairs from the UMNSRS dataset that appear in both the relatedness and similarity subsets. Each column shows the ranks obtained when ordering all 167 pairs from the subset according to the respective evaluation. The example pairs in the first block are the highest-scoring example pairs for each of the respective measures.The second block provides cases for which the human scores differ the most from each other as well as from the computed metrics. For instance, *overnutrition* and *malnutrition* are obviously related, but refer to different diagnostic circumstances. Interestingly, both AiTextML (label) and (word) evaluate the pair as closely related although the number of co-occurrences does also not indicate a strong connection. We suspect that the two words appear quite frequently in close vicinity, e.g. in enumerations. This imposes a proximity in the word embeddings space, which can then be transferred to the label space.

A similar case is *Xanax* vs. *Ativan*, two different drugs of the same active agent class for treatment of panic disorders, which AiTextML correctly assessed as highly similar although the co-occurrence patterns do not indicate it.

In contrast, a high number of co-occurrences seems to entail a high label embedding similarity. *Hypothyroidism* (a disorder in which the thyroid gland does not produce enough thyroid hormone) vs. *Synthroid* (a brand name for a synthetic thyroid hormone), for instance, is top-ranked according to co-occurrence, but not very similar according to word contexts.

In some cases, e.g. for *polydipsia* vs. *polyuria* and the already mentioned nutrition disorder pair, label and word-based similarities remain close. Yet, a rather high word co-occurrence frequency does not always imply closeness in the label embeddings space. For instance, *constipation* frequently comes with a temporal *anosmia*, the inability to perceive odor, which is hence ranked as 25^{th} pair according to our word embedding technique. Yet, this strong link is apparently insufficient to convince humans or our label embedding algorithm of a high similarity or relatedness.

6 Conclusion

Our experimental evaluation shows that the embeddings produced in our study correspond significantly better with human assessments of medical concept similarity and relatedness than previous semantic methods do. Moreover, our method is simpler to deploy than many traditional approaches. Once embeddings have been created, simple vector operations suffice to compute similarity scores. The original MeSH data no longer needs to be distributed with the tool. This suggests that our results are a simple and ef-

fective drop-in replacement to improve any systems relying on medical concept similarity measures. While our method depends on label annotations, these are often quite abundant, e.g. manually entered keywords for scientific publications or hashtags in social media. Our data can be freely accessed from `http://www.ke.tu-darmstadt.de/resources/medsim`.

Finally, vector representations encode various forms of semantic information that can be used for tasks beyond mere relatedness computation. An initial embedding layer can be used with neural network architectures or also to generate feature vectors for other machine learning methods. These machine learning algorithms can then make better predictions based on the semantics of the concepts rather than merely memorizing token identities.

Acknowledgements

This work has been supported by the German Institute for Educational Research (DIPF) under the Knowledge Discovery in Scientific Literature (KDSL) program, and the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources (AIPHES)" under grant No. GRK 1994/1.

Gerard de Melo's research is supported by China 973 Program Grants 2011CBA00300, 2011CBA00301, and NSFC Grants 61033001, 61361136003, 61550110504.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Chen, J. and de Melo, G. (2015). Semantic information extraction for improved word embeddings. In *Proceedings of the NAACL Workshop on Vector Space Modeling for NLP*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

de Melo, G. (2015). Wiktionary-based word embeddings. In *Proceedings of MT Summit XV*, pages 346–359. AMTA.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407.

Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING 1997)*.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning*, pages 1188–1196.

Liu, Y., McInnes, B. T., Pedersen, T., Melton-Meaux, G., and Pakhomov, S. (2012). Semantic Relatedness Study Using Second Order Co-occurrence Vectors Computed from Biomedical Corpora, UMLS and WordNet. In *Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium*, pages 363–372.

McInnes, B., Pedersen, T., and Pakhomov, S. (2009). UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*, San Fransico, CA, November.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Nam, J., Loza Mencía, E., and Fürnkranz, J. (2016). All-in text: Learning document, label, and word representations jointly. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.

Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., and Melton, G. B. (2010). Semantic similarity and relatedness between clinical terms: an experimental study. In *AMIA annual symposium proceedings*, volume 2010, pages 572–576.

Pedersen, T., Pakhomov, S. V. S., Patwardhan, S., and Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *J. of Biomedical Informatics*, 40(3):288–299, June.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543, October.

Pyysalo, S., Ginter, F., Moen, H., Salakoski, T., and Ananiadou, S. (2013). Distributional Semantics Resources for Biomedical Text Processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*, pages 39 – 43.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Schütze, H. (1993). Word space. In *Advances in Neural Information Processing Systems 5 (NIPS 1992)*.

Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL 2010*, pages 384–394.

Weston, J., Bengio, S., and Usunier, N. (2011). Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2764–2770. AAAI Press.

Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *Proceedings of ACL 1994*, pages 133–138.