# Sub-word information in pre-trained biomedical word representations: evaluation and hyper-parameter optimization

**Dieter Galea** **Ivan Laponogov** **Kirill Veselkov**
Department of Surgery and Cancer
Imperial College London
{d.galea14 | i.laponogov | kirill.veselkov04}@imperial.ac.uk

## Abstract

Word2vec embeddings are limited to computing vectors for in-vocabulary terms and do not take into account sub-word information. Character-based representations, such as fastText, mitigate such limitations. We optimize and compare these representations for the biomedical domain. fastText was found to consistently outperform word2vec in named entity recognition tasks for entities such as chemicals and genes. This is likely due to gained information from computed out-of-vocabulary term vectors, as well as the word compositionality of such entities. Contrastingly, performance varied on intrinsic datasets. Optimal hyper-parameters were intrinsic dataset-dependent, likely due to differences in term types distributions. This indicates embeddings should be chosen based on the task at hand. We therefore provide a number of optimized hyper-parameter sets and pre-trained word2vec and fastText models, available on https://github.com/dterg/bionlp-embed.

## 1 Introduction

word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) models are a popular choice for word embeddings, representing words by vectors for downstream natural language processing. Optimization of word2vec has been thoroughly investigated by Chiu et al. (2016a) for biomedical text. However, word2vec has two main limitations: i) out-of-vocabulary (OOV) terms cannot be represented, losing potentially useful information; and ii) training is based on co-occurrence of terms, not taking into account sub-word information. With new entities such as genetic variants, pathogens, chemicals and drugs, these limitations can be critical in biomedical NLP.

Sub-word information has played a critical role in improving NLP task performances and has predominantly depended on feature-engineering. More recently, character-based neural networks for tasks such as named entity recognition have been developed and evaluated on biomedical literature (Gridach, 2017). This has achieved state-of-the-art performances but is limited by the quantity of supervised training data.

Character-based representation models such as fastText (Bojanowski et al., 2017; Mikolov et al., 2018) and MIMICK (Pinter et al., 2017) exploit word compositionality to learn distributional embeddings, allowing to compute vectors for OOV words. Briefly, fastText uses a feed-forward architecture to learn n-gram and word embeddings, whereas MIMICK uses a Bi-LSTM architecture to learn character-based embeddings in the same space of another pre-trained embeddings, such as word-based word2vec.

Here we evaluate and optimize pre-trained character-based word representations with the fastText implementation for biomedical terms. To compare with word2vec models, we also optimize word2vec by extending the work by Chiu et al. (2016a). We report that fastText outperforms word2vec in all named entity recognition tasks of feature-rich entities such as chemicals and genes. However, in intrinsic evaluation, results and optimal hyper-parameters vary. This is likely due to different entity type distributions within the intrinsic standards. This indicates representations should be selected and optimized based on the task at hand and the entities of interest. We evaluate and provide optimized generalized fastText and word2vec models and models optimized on



*Proceedings of the BioNLP 2018 workshop*, pages 56–66
Melbourne, Australia, July 19, 2018. 

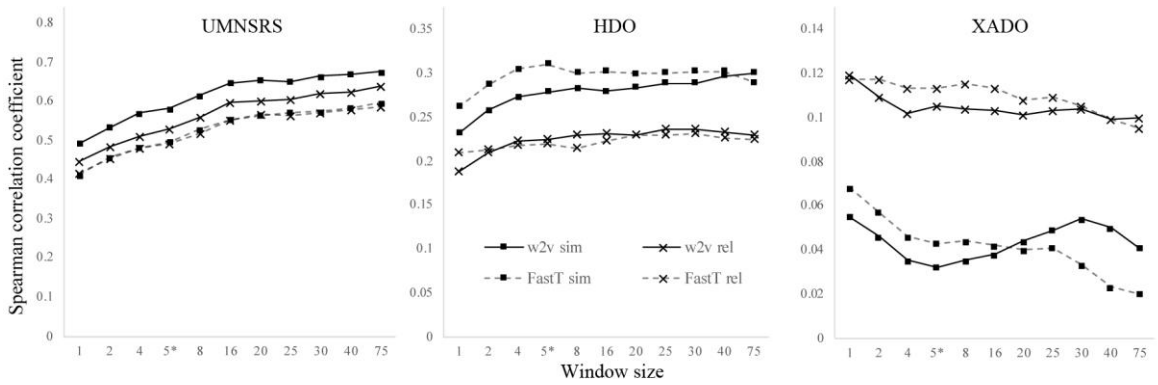


Figure 1: Intrinsic evaluation of window size in word2vec (w2v) and fastText (FastT) models on UMNSRS, HDO, and XADO datasets (Supp. Table 4).

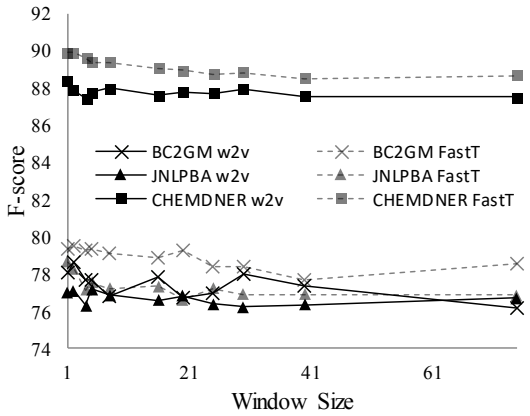


Figure 2: Extrinsic evaluation of window size in word2vec (w2v) and fastText (FastT) models on BC2GM, JNLPBA and CHEMDNER datasets (Supp. Table 5).

individual datasets, outperforming a number of current state-of-the-art embeddings.

## 2 Materials and Methods

### 2.1 Data and pre-processing

PubMed 2018 baseline abstracts and titles were parsed using PubMed parser (Achakulvisut and Acuna, 2016), each article was represented as a single line, and any new line characters within an article were replaced by a whitespace. Pre-processing was performed using the NLPre module (He and Chen, 2018). All upper-case sentences were lowered, de-dashed, parenthetical phrases identified, acronyms replaced with full term phrases (e.g. "Chronic obstructive pulmonary disease (COPD)" was changed to "Chronic pulmonary disease (Chronic_pulmonary_disease)), URLs removed, and single character tokens removed. Tokenization was carried out on whitespace. Punctuation was retained. This resulted in a training dataset of 3.4 billion tokens and a vocabulary of up to 19 million terms (Supp. Table 1).

### 2.2 Embeddings and hyper-parameters

Word embeddings were trained on the pre-processed PubMed articles using Skip-Gram word2vec and fastText implementations in gensim (Řehůřek and Sojka, 2010). As in Chiu et al. (2016a), we tested the effect of hyper-parameter selection on embedding performance for each hyper-parameter: negative sample size, sub-sampling rate, minimum word count, learning rate (alpha), dimensionality, and window size. Extended parameter ranges were tested for some hyper-parameters, such as window size. Additionally, we test the range of character n-grams for the fastText models, as originally performed for language models (Bojanowski et al., 2017). Due to the computational cost, especially since fastText models can be up to 7.2x slower to train compared to word2vec (Supp. Figure 1), we modify one hyper-parameter at a time, while keeping all other hyper-parameters constant. Performance was measured both intrinsically and extrinsically on a number of datasets.

### 2.3 Intrinsic Evaluation

Intrinsic evaluation of word embeddings is commonly performed by correlating the cosine similarity between term pairs, as determined by the trained embeddings, and a reference list. We use the manually curated UMNSRS covering disorders, symptoms, and drugs (Pakhomov et al., 2016), and compute graph-based similarity and relatedness using the human disease ontology graph (Schriml et al., 2012) (HDO) and the *Xenopus* anatomy and development ontology graph (Segerdell et al., 2008) (XADO). 1 million pairwise combinations of entities and ontologies were

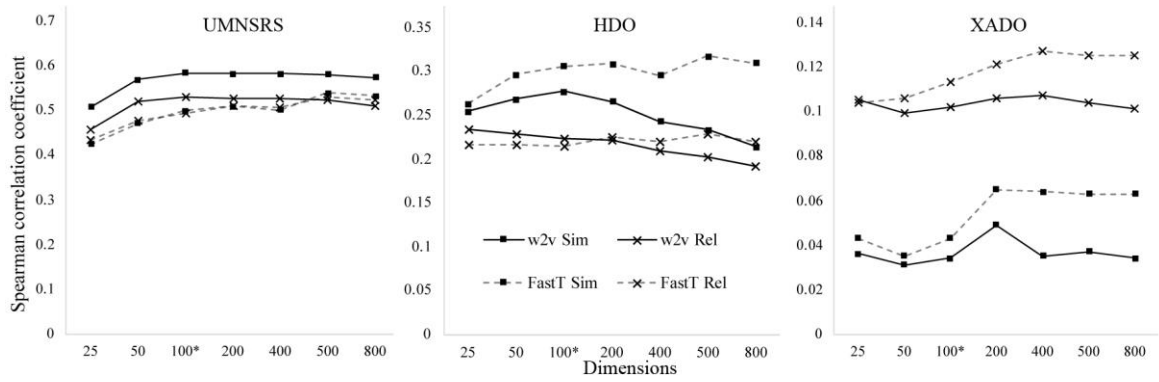


Figure 3: Intrinsic evaluation of dimension size in word2vec (w2v) and fastText (FastT) models on UMNSRS, HDO, and XADO datasets (Supp. Table 6).

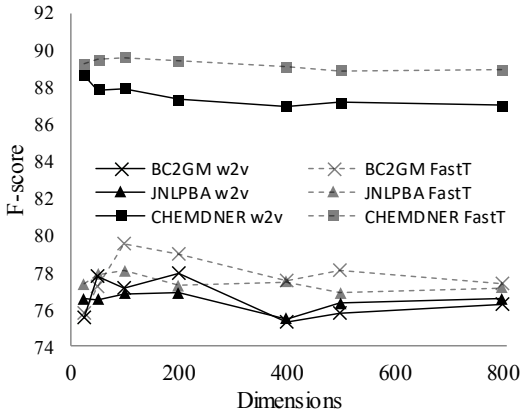


Figure 4: Extrinsic evaluation of dimension size in word2vec (w2v) and fastText (FastT) models on BC2GM, JNLPBA and CHEMDNER datasets (Supp. Table 7).

randomly computed from each graph and entities which did not map to the ontology map or were multi-token were not considered. Similarity between a pair of terms was computed using the Wu and Palmer (1994) similarity metric, and relatedness was determined by a simplified Lesk algorithm (1986). In the latter, token intersection (excluding stopwords) was calculated between definitions and normalized by the maximum definition length. Pairs which did not have definition statements for any of the terms were excluded.

As with UMNSRS, the computed similarity and relatedness scores were correlated with the cosine similarity determined by the embeddings models. As word2vec is not capable of representing OOV words, in literature pair terms which are not in vocabulary are commonly not considered for evaluation. To allow for comparison between the word2vec and fastText models, we represent OOV words as null vectors – as originally performed by Bojanowski et al. (2017). However, to determine the difference in performance of in-vocabulary word embeddings and OOV word embeddings, we measure correlation with only in-vocabulary terms, and with OOV terms pairs considered and null-imputed for word2vec.

### 2.4 Extrinsic evaluation

Intrinsic evaluation by itself may provide limited insights and may not represent the true downstream performance (Faruqui et al. 2016; Chiu et al., 2016b). Therefore, we perform extrinsic evaluation using 3 named entity recognition corpora: (i) the BioCreative II Gene Mention task corpus (BC2GM) (Smith et al., 2008) for genes; (ii) the JNLPBA corpus (Kim et al., 2004) annotating proteins, cell lines, cell types, DNA, and RNA; and (iii) the CHEMDNER corpus (Krallinger et al., 2015) which annotates drugs and chemicals, as made available from Luo et al. (2017). Each of these corpora are originally split into a train, development, and test sets – the same splits and sentence ordering were retained here.

The state-of-the-art BiLSTM-CRF neural network architecture (Lample et al., 2016), as implemented in the anago package, was used to train NER models and predict the development set of each corpus for each parameter. Accuracy was determined by the F-score. Each model was run for up to 10 epochs and the best accuracy on the development set was recorded.

### 2.5 Optimized Embeddings

Hyper-parameters achieving the highest performance for each extrinsic corpus and intrinsic standard were determined for word2vec and fastText. Corpus-specific word2vec and fastText models were trained with the set of optimal hyper-parameters for each corpus, as each corpus annotates different entity classes. For a generalized optimal model, we also trained embeddings on optimal hyper-parameters determined across all cor-

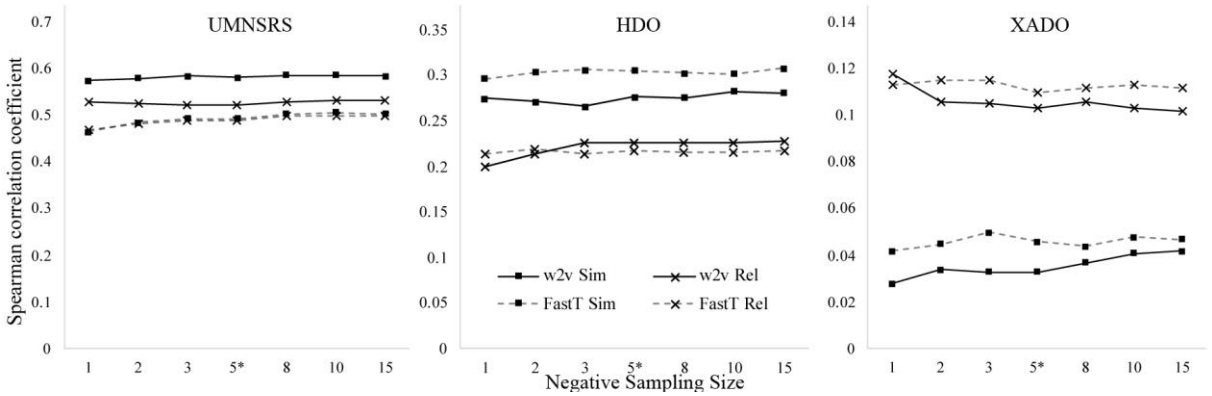


Figure 5: Intrinsic evaluation of negative sampling size in word2vec (w2v) and fastText (FastT) models on UMNSRS, HDO, and XADO datasets (Supp. Table 8).

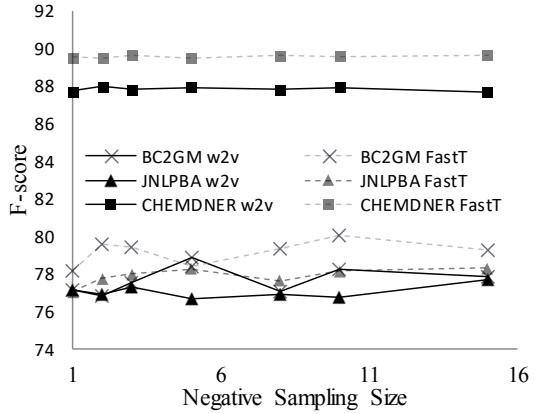


Figure 6: Extrinsic evaluation of negative sampling size in word2vec (w2v) and fastText (FastT) models on BC2GM, JNLPBA and CHEMDNER datasets (Supp. Table 9).

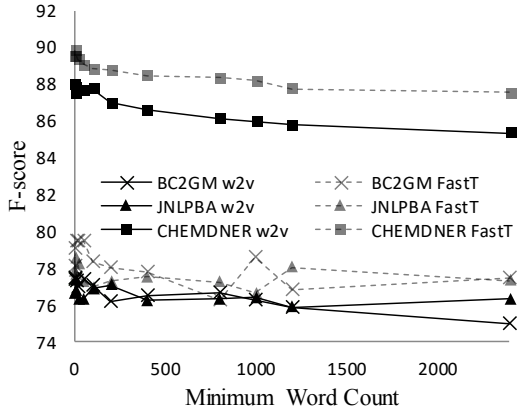


Figure 7: Extrinsic evaluation of minimum word count in word2vec (w2v) and fastText (FastT) models on BC2GM, JNLPBA and CHEMDNER datasets (Supp. Table 11).

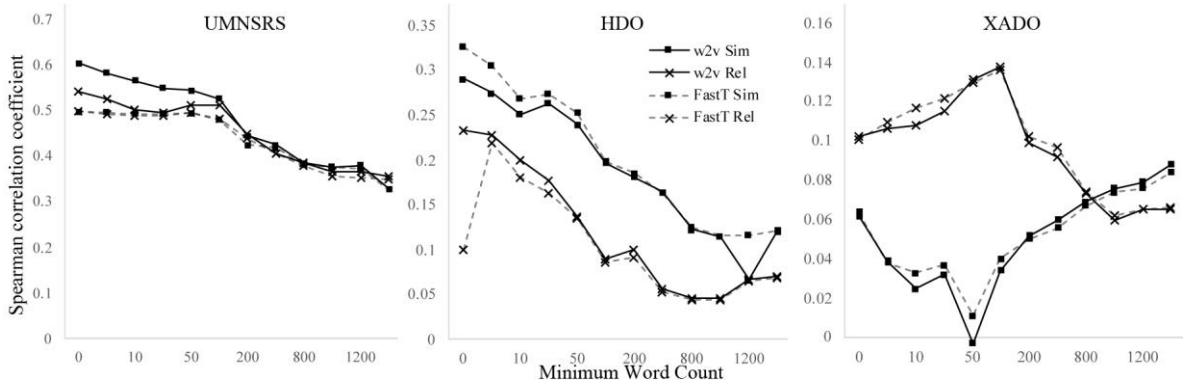


Figure 8: Intrinsic evaluation of the minimum word count in word2vec (w2v) and fastText (FastT) models on UMNSRS, HDO, and XADO datasets (Supp. Table 10).

pora and standards, as well as across intrinsic and extrinsic datasets separately. For the final extrinsic optimized evaluation, the test split was predicted.

## 3 Results and Discussion

### 3.1 General trends: word2vec hyper-parameter selection

Overall, intrinsic and extrinsic performance of word2vec models (Figure 1-12) obtained similar trends to Chiu et al. (2016a) for the same corpora/standards (i.e. UMNSRS, BC2GM, and JNLPBA), therefore we refer to Chiu et al. (2016a) for further discussion of these trends. Minor differences were recorded for minimum word count (Figure 7-8) and window size (Figure 1-2), where both UMNSRS similarity and relatedness

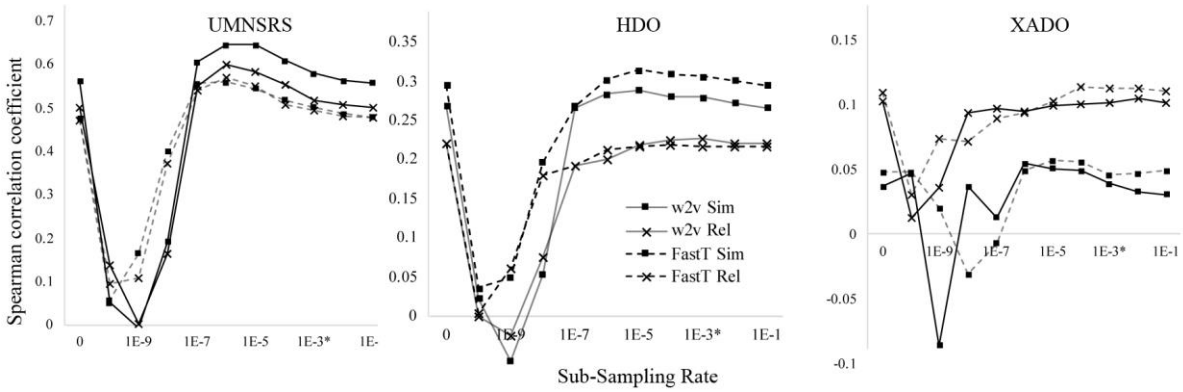


Figure 9: Intrinsic evaluation of sub-sampling rate in word2vec (w2v) and fastText (FastT) models on UMNSRS, HDO, and XADO datasets (Supp. Table 12).

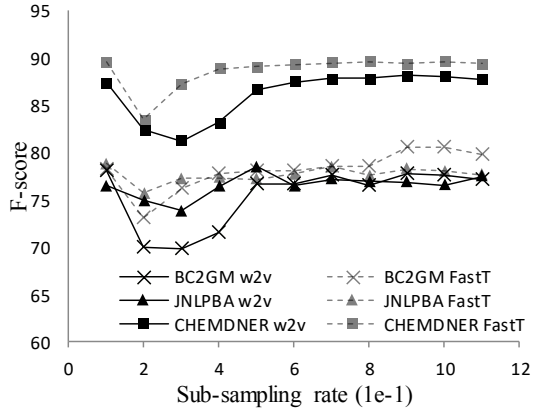


Figure 10: Extrinsic evaluation of sub-sampling rate in word2vec (w2v) and fastText (FastT) models on BC2GM, JNLPBA and CHEMDNER datasets (Supp. Table 13).

decreased with increasing minimum word count, whereas in Chiu et al. (2016a) this was only the case for relatedness.

In intrinsic evaluation of window size, particularly UMNSRS (Figure 1), performance consistently increased with increasing window size. This trend was also reported by Chiu et al. (2016a), where the maximum window size of 30 obtained the highest similarity and relatedness. We reasoned that abstracts generally concern a single topic, therefore predicted that increasing the window size to the average abstract length would capture more relevant information. This was indeed the case, obtaining 0.675 and 0.639 for UMNSRS similarity and relatedness respectively, compared to 0.627 and 0.584 similarity and relatedness respectively reported by Chiu et al. (2016a) for PubMed. As higher intrinsic performance was obtained in our results for similar window sizes, the difference in performance is also contributed to by an increase in the training data and different preprocessing.

In the case of extrinsic evaluation, the best performance was generally obtained with lower window size – a similar trend to that reported in Chiu et al. (2016a).

### 3.2 General trends: fastText hyper-parameter selection

Except for the character n-gram hyper-parameter, fastText models share the same hyper-parameters with word2vec models. Overall, similar trends in both intrinsic and extrinsic performance were obtained for word2vec and fastText embeddings (Figure 1-12). However, optimal parameters were not necessarily identical, as discussed below.

### 3.3 Comparison of representations – Intrinsic evaluation

While the overall performance trends with various hyper-parameters for fastText are similar to those obtained by word2vec, we report a number of notable differences.

When intrinsically evaluated with UMNSRS, word2vec representations consistently achieved higher similarity and relatedness compared to fastText for hyper-parameters such as: window size, dimensions and negative sampling, irrespective of the selected hyper-parameters. However, evaluating with HDO and XADO intrinsic datasets, results were more variable. fastText tended to perform similar to or outperform word2vec across negative sampling size, dimensions and window size hyper-parameter ranges.

Differences in performance between datasets may be a result of differences in: (i) number of OOV terms; (ii) rarity of terms; and (iii) term types. As UMNSRS is a manually curated reference list of term pairs with the vocabulary of multiple corpora, including PubMed Central, only up

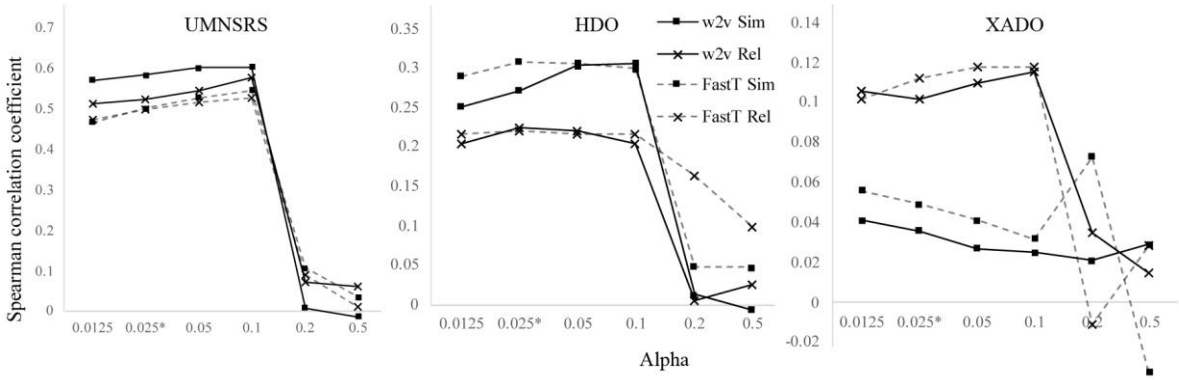


Figure 11: Intrinsic evaluation of the alpha hyper-parameter in word2vec (w2v) and fastText (FastT) models on UMNSRS, HDO, and XADO datasets (Supp. Table 14).

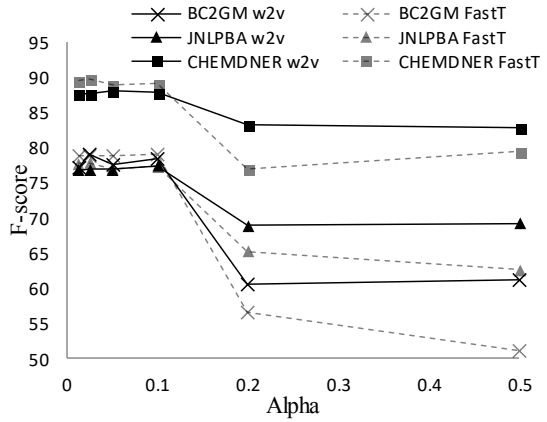


Figure 12: Extrinsic evaluation of alpha in word2vec (w2v) and fastText (FastT) models on BC2GM, JNLPBA and CHEMDNER datasets (Supp. Table 15).

to 9 total tokens were OOV (1.3%; Supp. Table 2). HDO contained up to 5% OOV terms. As OOV terms are represented by null vectors for word2vec models, a decrease in performance with increase in OOV terms is expected.

Skipping OOV term pairs from evaluation (rather than imputing) obtained similar performance trends across datasets, indicating that OOV is not the major contributing factor in such intrinsic performance differences. However, this may also imply that fastText degrades the performance for in-vocabulary terms of the UMNSRS dataset. Similar results were reported by the original authors when assessed on the English WS353 dataset (Bojanowski et al., 2017).

Despite terms being in-vocabulary, the frequency by which these occur in the training dataset may vary. This is indeed the case for UMNSRS and HDO, where UMNSRS has a median rank in-vocabulary frequency 4 times higher than HDO. This may indicate fastText provides better representations for rarer terms. XADO, however, has a median rank in-vocabulary frequency within 1.3 times of UMNSRS. This implies there are additional contributing factors to such performance differences, including potentially differences in the quality of the ontology graph.

As the intrinsic standards contain various entity classes, differences in representation models' performance (and optimal hyper-parameters) may be dependent on the distribution of entity types. fastText authors reported that fastText outperforms word2vec in languages like German, Arabic, Russian and in rare English words (Bojanowski et al., 2017). This indicates that word2vec and fastText's performance is dependent on the compositionality and word character features, and may therefore be expected to vary between biomedical entity classes.

Biomedical text generally contains terms such as chemicals, genes, proteins and cell-lines which are rich in features such as punctuation, special characters, digits, and mixed-case characters. Such orthographic features have been manually extracted in traditional machine learning methods, or more recently combined with word embeddings, and have been shown to have discriminating power in tasks such as named entity recognition (Galea et al., 2018).

### 3.4 Comparison of representations – Extrinsic evaluation

When performing named entity recognition as extrinsic evaluation of the word representations models, fastText consistently outperformed word2vec at any hyper-parameter value, and consistently across all 3 corpora (Figures 2,4,6,7,10,12). With 9-13% total OOV tokens, and 14-34% OOV entity tokens (Supp. Table 3, Supp. Fig. 3,4), this indicates the overall likely positive

| 1,2-dichloromethane | 1-(dimethylamino)-2-methyl-3,4-diphenylbutane-1,3-diol | ZNF560 |
|---|---|---|
| ***1,2-dichloroethane*** | 8-(N,N-***diethylamino***)oct***yl-3,4***,5-trimethoxybenzoate | ***ZNF5***80 |
| ***1,2-dichloro***benzene | 1,3-***dimethylam***yl***amin***e | ***ZNF5***45 |
| ***Di***bromo***chloromethane*** | 8-(***diethylamino***)oct***yl*** | ***ZNF5***82 |
| ***1,2-dichloro***prop***ane*** | 2-cyclohex***yl***-2-hydroxy-2-phen***yl***acetate | ***ZNF5***21 |
| water/***1,2-dichloroethane*** | ***diethylamin***e | SOX1 |

Table 1: Top 5 most similar words to a selection of out-of-vocabulary terms (two chemical systematic names and a protein symbol; top row). Sequences in bold indicate overlap with queried term.

contribution of gained information from computed OOV vectors.

In terms of the specific corpora, the largest performance difference was recorded for genes (BC2GM) and chemical names (CHEMDNER). As these two corpora only tag one entity type, entity variation is lower than JNLPBA which tags 5 entity classes and therefore this may contribute to the dissimilarities in performance difference between the corpora.

In addition to the rich and unique features, outperformance of fastText in extrinsic evaluation may also be attributed to the standardized nomenclature used in biomedical entities which provides additional within-token structure. For example, systematic chemical names follow the IUPAC nomenclature. Prefixes such as *mono*, *di*, and *tri* indicate number of identical substituents in a compound. Similarly, residual groups are represented by prefixes such as *methyl-* and *bromo-*. Additionally, the backbone structure of the molecule is assigned a suffix that indicates structure features (e.g. simple hydrocarbon molecules utilize suffixes to indicate number of single, double or more bonds, where *-ane* indicates single bonds, *-ene* double bonds, *-ynes* triple bonds etc).

With such structure, as fastText is a character-level model, for chemicals such as *1,2-dichloromethane*, most similar words include chemicals which share the substituents and their specific position, defined by the *1,2-dichloro-* prefix (Table 1). Therefore, fastText provides more structurally-similar chemicals, whereas word2vec would treat *1,2-dichloromethane* and *2-dichloromethane* as two completely different/unrelated terms (when excluding context or setting a small window size).

As chemicals can be synthesized and named, it is likely for very specific and big molecules such as *1-(dimethylamino)-2-methyl-3,4-diphenylbutane-1,3-diol* to be OOV. This is a great advantage of character-level embeddings which still enable computing a representation.

Given the highly standardized and structured nomenclature of chemicals, we briefly observed that fastText models are also able to recall structural analogs when performing analogy tasks. For example, methanol → methanal is an oxidation reaction where an alcohol is converted to an aldehyde, specifically the *-OH* group is converted to a *=O* group. Given ethanol and performing analogy task vector arithmetic, the aldehyde ethanal is returned. Similar results were observed for sulfuric_acid – sulfur + phosphorous, giving phosphoric_acid. Formal evaluation on analogy tasks is required to assess how character-based embeddings perform compared to word2vec.

Genes and proteins have full names as well as short symbolic identifiers which are usually acronymic abbreviations. These are less structured than chemical names, however, as the root portion of the symbols represents a gene family, this accounts for the similarity performance of character-based embeddings. *ZNF560* is an example of OOV protein that was assigned a vector close to ZNF* genes (Table 1) as well as *SOX1*. While *SOX1* does not share character n-grams with *ZNF560*, similarity was determined based on co-occurrence of *ZNF* genes and *SOX1* – genes which are associated with adenocarcinomas (Chang et al., 2015).

While the advantages of character-based similarity for OOV terms are clear, from intrinsic evaluation it appears that for some entities word2vec provides better embeddings. An example of this is when querying *phosphatidylinositol-4,5-bisphosphate* (Supp. Table 16). Whereas the top 5 most similar terms returned by fastText are orthographically, morphologically, and structurally similar, word2vec recalled *PIP2* and *PI(4,5)P2*. These are synonyms of the queried term hence more similar than *phosphatidylinositol-4-phosphate*, for example. A similar result was also observed for genetic variants (SNPs). While fastText returned *rs-* prefixed terms as most similar terms to the reference SNP identifier *rs2243250* (which refers to the SNP Interleukin 4 – 590C/T polymorphism), word2vec recalled terms *590C>T* and *590C/T*; the nucleotide polymorphism specified by the identifier itself (Supp

UMNSRS

| | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| 2 | 0.443 | 0.463 | 0.503 | 0.532 | 0.544 | 0.554 |
| | 0.427 | 0.458 | 0.497 | 0.507 | 0.512 | 0.516 |
| 3 | | 0.487 | 0.517 | 0.548 | 0.560 | 0.561 |
| | | 0.478 | 0.506 | 0.524 | 0.530 | 0.522 |
| 4 | | | 0.534 | 0.562 | 0.570 | 0.582 |
| | | | 0.523 | 0.539 | 0.533 | 0.540 |
| 5 | | | | 0.584 | 0.603 | 0.596 |
| | | | | 0.554 | **0.565** | 0.552 |
| 6 | | | | | **0.612** | 0.607 |
| | | | | | 0.556 | 0.549 |
| 7 | | | | | | 0.601 |
| | | | | | | 0.542 |

HDO

| | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| 2 | 0.219 | 0.282 | 0.295 | 0.296 | 0.302 | 0.302 |
| | 0.193 | 0.218 | 0.219 | 0.222 | 0.223 | 0.224 |
| 3 | | 0.298 | 0.307 | 0.307 | 0.312 | 0.312 |
| | | 0.213 | 0.217 | 0.215 | 0.226 | 0.225 |
| 4 | | | 0.313 | *0.318* | 0.316 | 0.315 |
| | | | 0.218 | 0.227 | 0.228 | 0.227 |
| 5 | | | | **0.320** | *0.319* | **0.320** |
| | | | | 0.230 | 0.226 | 0.228 |
| 6 | | | | | 0.317 | *0.319* |
| | | | | | 0.228 | **0.234** |
| 7 | | | | | | 0.314 |
| | | | | | | 0.231 |

XADO

| | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| 2 | 0.031 | 0.047 | 0.039 | 0.032 | 0.030 | 0.024 |
| | 0.105 | 0.106 | 0.106 | 0.113 | 0.114 | 0.114 |
| 3 | | **0.054** | 0.048 | 0.038 | 0.032 | 0.030 |
| | | 0.111 | 0.112 | **0.117** | **0.117** | **0.117** |
| 4 | | | 0.040 | 0.036 | 0.035 | 0.030 |
| | | | 0.110 | 0.111 | 0.113 | 0.109 |
| 5 | | | | 0.034 | 0.031 | 0.029 |
| | | | | 0.112 | 0.108 | 0.109 |
| 6 | | | | | 0.037 | 0.035 |
| | | | | | 0.110 | 0.108 |
| 7 | | | | | | 0.033 |
| | | | | | | 0.102 |

BC2GM

| | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| 2 | 78.96 | *79.72* | *79.78* | *79.91* | *79.71* | **80.26** |
| 3 | | *79.69* | 78.88 | 78.77 | *79.90* | *79.91* |
| 4 | | | 78.94 | 78.42 | 78.91 | 79.05 |
| 5 | | | | 77.12 | 78.67 | 77.45 |
| 6 | | | | | 77.77 | 77.97 |
| 7 | | | | | | 77.73 |

JNLPBA

| | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| 2 | *78.20* | 77.76 | 77.99 | 77.89 | 77.96 | 77.96 |
| 3 | | 77.83 | 77.86 | 77.67 | **78.46** | *78.30* |
| 4 | | | 77.58 | 76.91 | 78.00 | 77.58 |
| 5 | | | | 76.72 | 78.04 | 76.92 |
| 6 | | | | | 77.82 | 78.06 |
| 7 | | | | | | 77.83 |

CHEMDNER

| | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| 2 | 89.14 | 89.48 | *89.66* | *89.72* | 89.58 | 89.46 |
| 3 | | 89.48 | *89.67* | *89.67* | **89.75** | *89.62* |
| 4 | | | 89.28 | 89.37 | 89.22 | 89.32 |
| 5 | | | | 89.22 | 89.13 | 89.10 |
| 6 | | | | | 89.03 | 88.81 |
| 7 | | | | | | 88.60 |

Table 2: Intrinsic (UMNSRS, HDO, XADO; upper row = similarity, lower row = relatedness) and extrinsic (BC2GM, JNLPBA, CHEMDNER) evaluation of the effect of character n-gram ranges on performance. Highest absolute accuracy is indicated in bold and accuracies within the standard error of the highest accuracy is italicized.

| | | UMNSRS | | HDO | | XADO | | BC2GM | JNLPBA | CHEMDNER |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Sim | Rel | Sim | Rel | Sim | Rel | | | |
| int | w2v | **0.726** | **0.690** | 0.314 | 0.237 | **0.095** | 0.077 | 76.43 | 71.84 | 87.83 |
| | FastT | 0.694 | 0.659 | **0.330** | **0.243** | 0.074 | 0.093 | 76.48 | 72.47 | 88.89 |
| ex | w2v | 0.506 | 0.469 | 0.252 | 0.184 | 0.024 | **0.120** | 77.13 | 73.61 | 88.93 |
| | FastT | 0.479 | 0.446 | 0.283 | 0.221 | 0.054 | 0.116 | **79.63** | **74.29** | **90.14** |

Table 3. Intrinsic and extrinsic performance for word2vec and fastText models optimized on optimum hyper-parameters from intrinsic (int) and extrinsic (ex) datasets (Supp. Table 27).

Table 19).

Additional examples comparing word2vec and fastText's most similar terms for chemicals, genes and diseases are provided in Supp. Tables 18-22.

From the quantitative results and the above qualitative examples, we observe a trade-off between character sequence similarity and context. The importance of which depends on the entity types – just as different languages benefit differently from word2vec and fastText models (Bojanowski et al., 2017).

### 3.5 Effect of n-grams size

Intrinsic evaluation shows high variability in the range of n-grams between the different standards (Table 2 & Supp. Table 25). UMNSRS achieves the highest performance (in terms of similarity) with 6-7 n-grams, whereas XADO achieves best results with 3-4 n-grams, and HDO achieves equal performance with ranges: 5-{6,7,8}, 4-6 and 6-8. This indicates the heterogeneity of the terms, both within the reference standards for HDO and XADO, and between standards. This further backs up the difference between the representation models due to entity type differences.

Contrastingly, extrinsic evaluation showed high consistency in n-gram ranges, with all corpora recording highest performance for the ranges 3-7 and 3-8. Within standard error (Supp. Table 23, 24), high performance was also obtained for ranges with lower limit of 2 and 3. Such ranges indicate that both short and long n-grams provide relevant information, complying with the previous discussion and examples for gene nomenclature and chemical naming conventions.

### 3.6 Optimized Models

Word embeddings trained on individual reference standards' optimal hyper-parameters (Supp. Table 25) achieved 0.733/0.686 similarity/relatedness with word2vec for UMNSRS (Supp. Table 26). This exceeds 0.652/0.601 reported by Chiu et al. (2016a), and the more recent 0.681/0.635 by Yu et al. (2017) achieved by retrofitting representations with knowledgebases, but not 0.75/0.73 by MeSH2Vec using prior knowledge (Jha et al., 2017). We expect further improvement to our models by retrofitting and augmenting prior knowledge.

Corpus-optimized fastText embeddings outperformed word2vec across all extrinsic corpora, recording: 79.33%, 73.30% and 90.54% for BC2GM, JNLPBA, and CHEMDNER (Supp. Table 26). This outperforms Chiu et al. (2016a), Pyysalo et al. (2013) and Kosmopoulos et al. (2015), although differences are also due to differ-

ent NER architectures used. However, our 90.54% CHEMDNER performance outperforms 89.28% using similar architectures and is close to the 90.84% achieved for attention-based architectures (Luo et al., 2017) - the best performance reported in literature to date.

Optimizing word2vec and fastText representations across all corpora and standards (Supp. Table 28) decreased the performance difference in NER between word2vec and fastText. This is due to the differences in the optimal hyper-parameters between intrinsic and extrinsic data (Supp. Table 29). Based on these differences, and as it had been shown that intrinsic results are not reflective of extrinsic performance (Chiu et al. 2016b), we generated separate word2vec and fastText models optimized on intrinsic and extrinsic datasets separately (Table 3). Again, fastText outperforms word2vec in all NER tasks but only outperforms word2vec for the HDO intrinsic dataset, possibly due to similarity implied from disease suffixes captured by n-grams.

## 4 Conclusion and future directions

We show that fastText consistently outperforms word2vec in named entity recognition of entities such as chemicals and genes. This is likely to be contributed to by the ability of character-based representations to compute vectors for OOV, and due to the highly structured, standardized and feature-rich nature of such entities.

Intrinsic evaluation indicated that the optimal hyper-parameter set, and hence optimal performance, is highly dataset-dependent. While number of OOV terms and rarity of in-vocabulary terms may contribute to such differences, further investigation is required to determine how the different entity types within the corpora are affected. Similarly, for named entity recognition, investigating the performance differences for each entity class would provide a more fine-grained insight into which classes benefit mostly from fastText, and why.

Empirically, we observed a trade-off between character sequence similarity and context in word2vec and fastText models. It would be interesting to assess how embedding models such as MIMICK, where the word2vec space can be preserved while still being able to generate character-based vectors for OOV terms, compare.

## Acknowledgements

We acknowledge financial support by BBSRC (BB/L020858/1), Imperial College Stratified Medicine Graduate Training Programme in Systems Medicine and Spectroscopic Profiling (STRATiGRAD), and Waters corporation.

## References

Titipat Achakulvisut, and Daniel E. Acuna. 2016. Pubmed Parser. https://doi.org/10.5281/zenodo.159504

Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, pages 135-146. http://aclweb.org/anthology/Q17-1010

Cheng-Chang Chang, Yu-Che Ou, Kung-Liahng Wang, Ting-Chang Chang, Ya-Min Cheng, Chi-Hau Chen, et al. 2015. Triage of Atypical Glandular Cell by SOX1 and POU4F3 methylation: A Taiwanese gynecologic oncology group (TGOG) study. *PLoS ONE* 10(6): e0128705. https://doi.org/10.1371/journal.pone.0128705

Billy Chiu, Gamal Crichton, Anna Korhonen and Sampo Pyysalo. How to train good word embeddings for biomedical NLP. 2016a. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166-174. https://doi.org/10.18653/v1/W16-2922

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016b. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluation Vector Space Representations for NLP*, pages 1-6. https://doi.org/10.18653/v1/W16-2501

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP,* pages 30-35. http://www.aclweb.org/anthology/W16-2506

Dieter Galea, Ivan Laponogov and Kirill Veselkov. 2018. Exploiting and assessing multi-source data for supervised biomedical named entity recognition. *Bioinformatics*, bty152. https://doi.org/10.1093/bioinformatics/bty152

Mourad Gridach. 2017. Character-level neural network for biomedical named entity recognition. *Journal*

*of Biomedical Informatics.* 70, pages 85-91. https://doi.org/10.1016/j.jbi.2017.05.002

Jiangen He and Chaomei Chen. 2018. Predictive effects of novelty measured by temporal embeddings on the growth of scientific literature. *Frontiers in Research Metrics and Analytics*, 3, 9. https://doi.org/10.3389/frma.2018.00009

Kishlay Jha, Guangxu Xun, Vishrawas Gopalakrishnan, and Aidong Zhang. 2017. Augmenting word embeddings through external knowledge-base for biomedical application. *IEEE International Conference on Big Data*, pages 1965-1974. https://doi.org/10.1109/BigData.2017.8258142

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746-1751. https://doi.org/10.3115/v1/D14-1181

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of JNLPBA*, pages 70–75. http://www.aclweb.org/anthology/W04-1213

Aris Kosmopoulos, Ion Androutsopoulos, and Georgios Paliouras. 2015. Biomedical semantic indexing using dense word vectors in BioASQ. *Journal of Biomedical Semantics.*

Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal and Alfonso Valencia. 2015. CHEMDNER: The drugs and chemical names extraction challenge. *Journal of Cheminformatics*, 7(Suppl 1):S1. https://doi.org/10.1186/1758-2946-7-S1-S1

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT,* pages 260-270. https://doi.org/10.18653/v1/N16-1030

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *In Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26. https://doi.org/10.1145/318723.318728

Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. 2017. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, btx761. https://doi.org/10.1093/bioinformatics/btx761

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111-3119

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. 2018. In *Proceedings of the International Conference on Language Resources and Evaluation.*

Serguei V.S. Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B. Melton. 2016. Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*, 32:23, pages 3635-3644. https://doi.org/10.1093/bioinformatics/btw529

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532-1543. https://doi.org/10.3115/v1/D14-1162

Yuval Pinter, Robert Guthrie, and Jacon Eisenstein. Mimicking word embeddings using subword RNNs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102-112. https://doi.org/10.18653/v1/D17-1010

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of LBM.*

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. 2010. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*, pages 46-50. https://doi.org/10.13140/2.1.2393.1847

Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng and Warren Alden Kibbe. 2012. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Research,* 40: D940-D945. https://doi.org/10.1093/nar/gkr972

Erik Segerdell, Jeff B. Bowes, Nicolas Pollet and Peter D. Vize. 2008. An ontology for *Xenopus* anatomy and development. *BMC Developmental Biology.* 8:92. https://doi.org/10.1186/1471-213X-8-92

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of biocreative

ii gene mention recognition. *Genome biology*, 9(Suppl 2):1–19. https://doi.org/10.1186/gb-2008-9-s2-s2

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Meeting of Association of Computational Linguistics*, pages 33–138. https://doi.org/10.3115/981732.981751

Zhiguo Yu, Byron C. Wallace, Todd Johnson, and Trevor Cohen. 2017. Retrofitting concept vector representations of medical concepts to improve estimates of semantic similarity and relatedness. In *Proceedings of the 16th World Congress on Medical and Health Informatics*