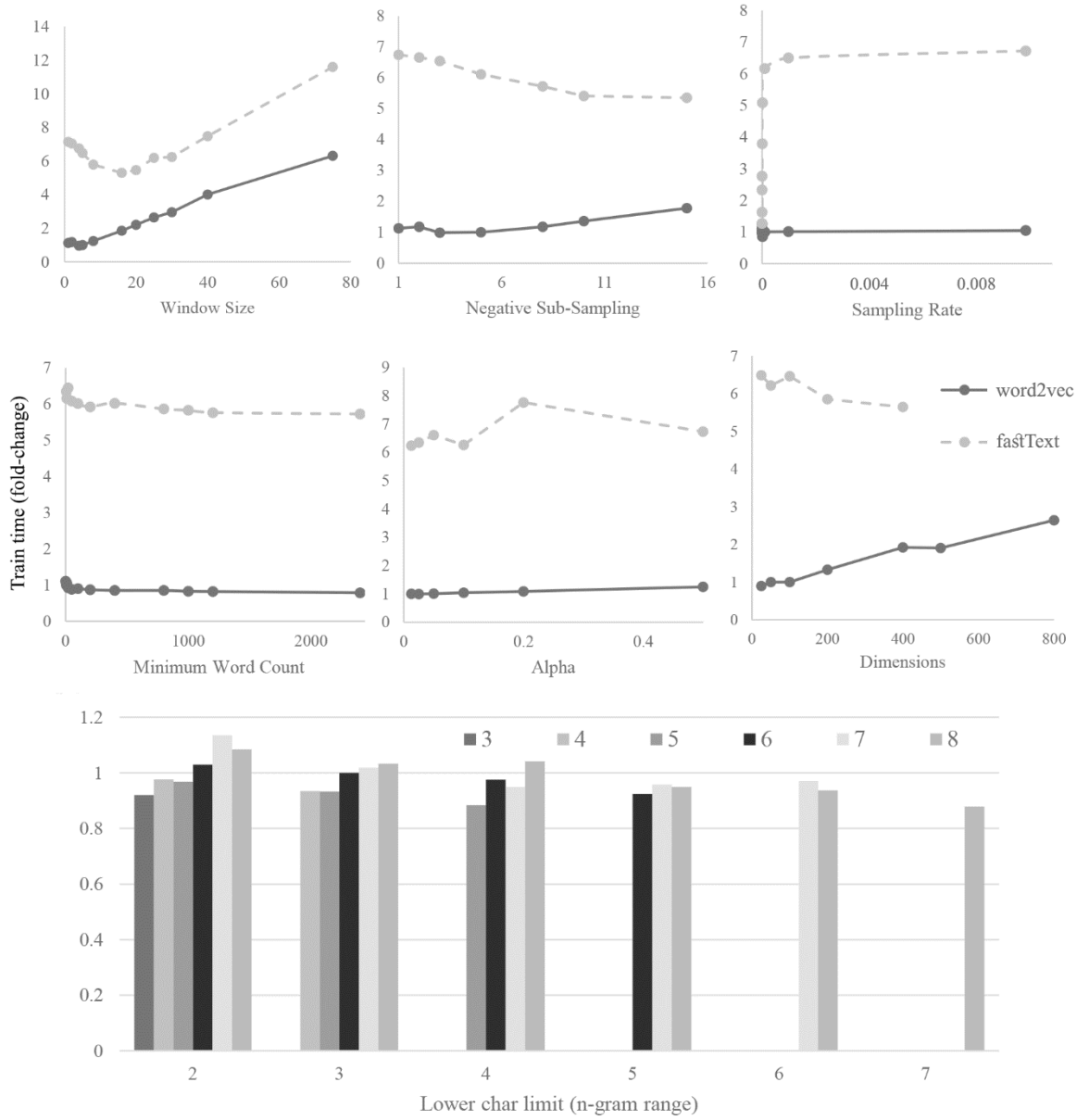
Supplementary Data



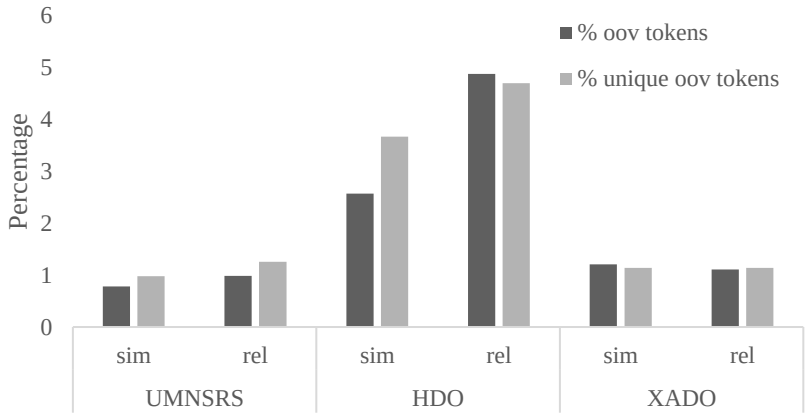
Supplementary Figure 1. Computational time for training Skip-Gram word2vec and fastText word embeddings at different hyper-parameter settings for: window size, negative sampling size, sub-sampling rate, minimum word count, alpha, and dimensionality. Plots are normalized with respect to the train time of word2vec when using default parameters (70-minute train time), and the default 3-6 character range in fastText (490-minute train time). Gensim's implementation of word2vec and fastText was used.

Supplementary Table 1. Total number of PubMed tokens and unique tokens used to train the embedding models at 3 different word frequency thresholds: all tokens, minimum word frequency of 5 and 10.

Minimum word frequency threshold	0	5	10
Tokens	3,435,773,079	3,412,644,449	3,402,300,795
Vocabulary	19,099,369	3,410,473	1,806,181

Supplementary Table 2. Intrinsic similarity and relatedness corpora (UMNSRS, HDO and XADO) token statistics. Total number of tokens, total out-of-vocabulary (OOV) tokens, number of unique tokens, and unique OOV tokens for all tokens. OOV statistics are based on default minimum word count hyper-parameter of 5.

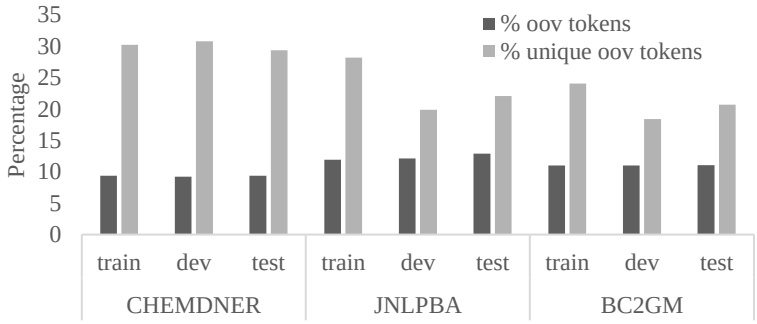
	UMNSRS		**HDO**		**XADO**	
	sim	**rel**	**sim**	**rel**	**sim**	**rel**
Total tokens	898	916	79978	61182	306839	334633
Unique tokens	306	318	873	511	264	264
Total OOV tokens	7	9	2057	2984	3704	3704
Unique OOV tokens	3	4	32	24	3	3



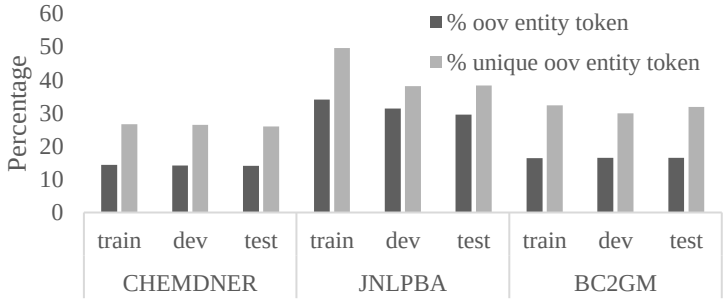
Supplementary Figure 2. Percentage total and unique out-of-vocabulary (OOV) tokens for each intrinsic corpus (UMNSRS, HDO, and XADO). Percentage OOV tokens is with respect to "Total tokens", whereas percentage unique OOV tokens is with respect to "Unique tokens" (Supplementary Table 2).

Supplementary Table 3. Extrinsic corpora (CHEMDNER, JNLPBA, BC2GM) token statistics. Total number of tokens, total out-of-vocabulary (OOV) tokens, number of unique tokens, and unique OOV tokens for all tokens and entity tokens. OOV statistics are based on default minimum word count hyper-parameter of 5.

	CHEMDNER			JNLPBA			BC2GM		
	train	**dev**	**test**	**train**	**dev**	**test**	**train**	**dev**	**test**
Total tokens	644195	639299	551617	446890	47661	101443	355405	71042	143465
Unique tokens	40996	40867	37245	20712	5679	9624	29774	11047	17344
Total OOV tokens	60293	58845	51586	53360	5767	13066	39173	7805	15883
Unique OOV tokens	12390	12568	10932	5841	1129	2124	7166	2033	3590
Total OOV entity tokens	8662	8326	7235	18153	1807	3850	6426	1287	2620
Unique OOV entity tokens	3299	3318	2837	2899	430	813	2312	608	1143



Supplementary Figure 3. Percentage total and unique out-of-vocabulary (OOV) tokens for each extrinsic corpus (CHEMDNER, JNLPBA, and BC2GM). Percentage OOV tokens is with respect to "Total tokens", whereas percentage unique OOV tokens is with respect to "Unique tokens" (Supplementary Table 3).



Supplementary Figure 4. Percentage total and unique out-of-vocabulary (OOV) entity tokens for each extrinsic corpus (CHEMDNER, JNLPBA, and BC2GM). Any token that was labeled as part of an entity in the original corpus (using the IOB scheme), was considered. Percentage OOV entity tokens is with respect to “Total OOV entity tokens”, whereas percentage unique OOV entity tokens is with respect to “Unique OOV entity tokens” (Supplementary Table 3).

Supplementary Table 4. **Effect of window size on intrinsic performance**. Intrinsic evaluation of word2vec (w2v) and fastText (FastT) word embedding models with various values for the "window size" hyper-parameter. Similarity and relatedness is represented by the Spearman correlation coefficient for the cosine similarity of term pairs and the reference standards. Out-of-vocabulary terms were considered and represented by a null vector for word2vec models. Asterisk (*) indicates default parameter value. Highest performance for each model and each standard is in **bold**.

Win	UMNSRS				HDO				XADO			
	w2v		FastT		w2v		FastT		w2v		FastT	
	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel
1	0.492	0.446	0.411	0.415	0.232	0.188	0.262	0.210	**0.055**	**0.119**	**0.068**	**0.117**
2	0.533	0.485	0.457	0.454	0.258	0.210	0.287	0.213	0.046	0.109	0.057	**0.117**
4	0.570	0.512	0.481	0.479	0.273	0.223	0.305	0.218	0.035	0.102	0.046	0.113
5*	0.580	0.528	0.497	0.491	0.279	0.224	**0.311**	0.220	0.032	0.105	0.043	0.113
8	0.614	0.558	0.528	0.517	0.283	0.229	0.301	0.214	0.035	0.104	0.044	0.115
16	0.647	0.596	0.552	0.552	0.280	0.231	0.302	0.223	0.038	0.103	0.042	0.113
20	0.654	0.601	0.563	0.565	0.284	0.230	0.300	0.230	0.044	0.101	0.040	0.108
25	0.651	0.603	0.571	0.564	0.289	0.236	0.301	0.229	0.049	0.103	0.041	0.109
30	0.663	0.618	0.574	0.571	0.289	**0.237**	0.302	**0.232**	0.054	0.104	0.033	0.105
40	0.670	0.624	0.583	0.578	0.297	0.233	0.302	0.227	0.050	0.099	0.023	0.099
75	**0.675**	**0.639**	**0.596**	**0.586**	**0.301**	0.229	0.290	0.225	0.041	0.100	0.020	0.095

Supplementary Table 5. **Effect of window size on extrinsic performance.** Extrinsic evaluation of word2vec (w2v) and fastText (FastT) word embedding models with various values for the "window size" hyper-parameter". Accuracies represent F-score for named entity recognition of 3 corpora: BC2GM, JNLPBA, and CHEMDNER. Asterisk (*) indicates default parameter value. Highest performance for each model and each corpus is in **bold**.

Win	BC2GM		JNLPBA		CHEMDNER	
	w2v	FastT	w2v	FastT	w2v	FastT
1	78.10	79.37	76.99	**78.64**	**88.34**	**89.89**
2	**78.61**	**79.51**	77.09	78.29	87.89	**89.89**
4	77.68	79.26	76.29	77.14	87.45	89.62
5*	77.64	79.31	**77.18**	77.52	87.74	89.38
8	76.81	79.14	76.84	77.22	87.98	89.40
16	77.84	78.88	76.58	77.33	87.60	89.09
20	76.70	79.25	76.82	76.61	87.80	88.96
25	76.93	78.40	76.40	77.23	87.70	88.75
30	77.97	78.42	76.21	76.91	87.95	88.82
40	77.33	77.66	76.35	76.88	87.55	88.50
75	76.19	78.58	76.69	76.86	87.51	88.67

Supplementary Table 6. **Effect of dimensionality on intrinsic performance.** Intrinsic evaluation of word2vec (w2v) and fastText (FastT) word embedding models with various values for the "dimension" hyper-parameter. Similarity and relatedness is represented by the Spearman correlation coefficient for the cosine similarity of term pairs and the 3 reference standards. Out-of-vocabulary terms were considered and represented by a null vector for word2vec models. Asterisk (*) indicates default parameter value. Highest performance for each model and each standard is in bold.

Dim	UMNSRS				HDO				XADO			
	w2v		FastT		w2v		FastT		w2v		FastT	
	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel
25	0.507	0.457	0.424	0.433	0.254	**0.234**	0.263	0.216	0.036	0.105	0.043	0.104
50	0.568	0.519	0.471	0.477	0.268	0.229	0.296	0.216	0.031	0.099	0.035	0.106
100*	**0.584**	**0.528**	0.498	0.493	**0.277**	0.223	0.306	0.215	0.034	0.102	0.043	0.113
200	0.581	0.526	0.509	0.508	0.266	0.222	0.308	0.225	**0.049**	0.106	**0.065**	0.121
400	0.582	0.527	0.500	0.505	0.243	0.210	0.295	0.219	0.035	**0.107**	0.064	**0.127**
500	0.580	0.522	**0.538**	**0.531**	0.233	0.203	**0.317**	**0.228**	0.037	0.104	0.063	0.125
800	0.574	0.510	0.532	0.523	0.214	0.192	0.309	0.220	0.034	0.101	0.063	0.125

Supplementary Table 7. **Effect of dimensionality on extrinsic performance.** Extrinsic evaluation of word2vec (w2v) and fastText (FastT) word embedding models with various values for the "dimension" hyper-parameter. Accuracies represent F-score for named entity recognition. Asterisk (*) indicates default parameter value. Highest performance for each model and each corpus is in bold.

Dim	BC2GM		JNLPBA		CHEMDNER	
	w2v	FastT	w2v	FastT	w2v	FastT
25	75.57	75.77	76.56	77.37	88.63	89.26
50	77.80	77.23	76.54	77.88	87.87	89.49
100*	77.12	**79.50**	76.83	**78.10**	**87.94**	**89.60**
200	**77.91**	79.01	**76.88**	77.28	87.34	89.42
400	75.28	77.56	75.50	77.50	86.99	89.07
500	75.76	78.07	76.31	76.90	87.18	88.89
800	76.23	77.41	76.57	77.17	87.02	88.91

Supplementary Table 8. **Effect of negative sampling on intrinsic performance.** Intrinsic evaluation of word2vec (w2v) and fastText (FastT) word embedding models with various values for the "negative sub-sampling" hyper-parameter. Similarity and relatedness is represented by the Spearman correlation coefficient for the cosine similarity of term pairs and the 3 reference standards. Out-of-vocabulary terms were considered and represented by a null vector for word2vec models. Asterisk (*) indicates default parameter value. Highest performance for each model and each standard is in **bold**.

Neg	UMNSRS				HDO				XADO			
	w2v		FastT		w2v		FastT		w2v		FastT	
	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel
1	0.575	0.529	0.466	0.471	0.275	0.200	0.297	0.215	0.028	**0.118**	0.042	0.113
2	0.579	0.527	0.485	0.483	0.272	0.214	0.304	**0.220**	0.034	0.106	0.045	**0.115**
3	0.584	0.524	0.494	0.488	0.266	0.226	0.307	0.214	0.033	0.105	**0.050**	**0.115**
5*	0.582	0.524	0.493	0.490	0.277	0.226	0.306	0.218	0.033	0.103	0.046	0.110
8	**0.587**	0.528	0.504	**0.499**	0.276	0.226	0.303	0.216	0.037	0.106	0.044	0.112
10	**0.587**	0.532	**0.507**	0.498	**0.283**	0.226	0.302	0.216	0.041	0.103	0.048	0.113
15	0.585	**0.533**	0.504	0.498	0.281	**0.229**	**0.308**	0.217	**0.042**	0.102	0.047	0.112

Supplementary Table 9. **Effect of negative sampling on extrinsic performance.** Extrinsic evaluation of word2vec (w2v) and fastText (FastT) word embedding models with various values for the "negative sub-sampling" hyper-parameter". Accuracies represent F-score for named entity recognition of 3 corpora: BC2GM, JNLPBA, and CHEMDNER. Asterisk (*) indicates default parameter value. Highest performance for each model and each corpus is in bold.

Neg	BC2GM		JNLPBA		CHEMDNER	
	w2v	FastT	w2v	FastT	w2v	FastT
1	77.11	78.10	77.09	77.00	87.74	89.53
2	76.78	79.52	76.90	77.71	**87.96**	89.50
3	77.48	79.40	77.30	77.96	87.79	89.64
5*	**78.87**	78.36	76.67	78.23	87.94	89.50
8	77.04	79.35	76.87	77.60	87.81	89.63
10	78.23	**80.00**	76.70	78.10	87.91	89.58
15	77.85	79.22	**77.68**	**78.26**	87.70	**89.66**

Supplementary Table 10. **Effect of minimum word count on intrinsic performance.** Intrinsic evaluation of word2vec (w2v) and fastText (FastT) word embedding models with various values for the "minimum word count" hyper-parameter. Similarity and relatedness is represented by the Spearman correlation coefficient for the cosine similarity of term pairs and the 3 reference standards. Out-of-vocabulary terms were considered and represented by a null vector for word2vec models. Asterisk (*) indicates default parameter value. Highest performance for each model and each standard is in bold.

Min-count	UMNSRS				HDO				XADO			
	w2v		FastT		w2v		FastT		w2v		FastT	
	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel
0	**0.604**	**0.542**	**0.499**	**0.498**	**0.291**	**0.234**	**0.327**	**0.101**	**0.062**	0.102	0.064	0.101
5*	0.583	0.526	0.496	0.493	0.275	0.229	0.306	0.220	0.039	0.106	0.038	0.110
10	0.567	0.504	0.494	0.488	0.251	0.200	0.269	0.181	0.025	0.108	0.033	0.117
20	0.550	0.497	0.494	0.490	0.264	0.177	0.274	0.164	0.032	0.115	0.037	0.122
50	0.545	0.512	0.495	0.496	0.240	0.138	0.253	0.135	-0.003	0.131	0.011	0.130
100	0.527	0.514	0.481	0.483	0.197	0.090	0.199	0.086	0.034	**0.138**	0.040	**0.136**
200	0.445	0.450	0.425	0.438	0.181	0.100	0.186	0.092	0.052	0.099	0.050	0.102
400	0.426	0.407	0.418	0.406	0.165	0.056	0.164	0.054	0.060	0.092	0.056	0.097
800	0.386	0.388	0.383	0.380	0.123	0.046	0.125	0.045	0.069	0.073	0.067	0.074
1000	0.378	0.365	0.376	0.358	0.115	0.047	0.116	0.044	0.076	0.060	0.074	0.062
1200	0.380	0.366	0.372	0.354	0.067	0.067	0.117	0.065	0.079	0.065	0.076	0.065
2400	0.329	0.356	0.327	0.349	0.122	0.070	0.121	0.069	0.088	0.065	**0.084**	0.066

Supplementary Table 11. **Effect of minimum word count on extrinsic performance.** Extrinsic evaluation of word2vec (w2v) and fastText (FastT) word embedding models with various values for the "minimum word count" hyper-parameter. Accuracies represent F-score for named entity recognition of 3 corpora: BC2GM, JNLPBA, and CHEMDNER. Asterisk (*) indicates default parameter value. Highest performance for each model and each corpus is in bold.

Min-count	BC2GM		JNLPBA		CHEMDNER	
	w2v	FastT	w2v	FastT	w2v	FastT
0	77.40	78.10	76.64	77.80	**87.99**	89.54
5*	**77.50**	79.07	76.80	**78.59**	87.84	**89.83**
10	77.34	79.46	**77.34**	77.64	87.55	89.56
20	76.97	79.45	76.35	78.22	87.71	89.35
50	77.36	**79.49**	76.35	77.29	87.67	89.08
100	77.04	78.33	76.89	76.96	87.73	88.81
200	76.20	78.06	77.10	77.31	86.97	88.76
400	76.52	77.81	76.25	77.52	86.56	88.45
800	76.68	76.31	76.33	77.27	86.13	88.33
1000	76.25	78.62	76.43	76.67	85.97	88.19
1200	75.87	76.8	75.90	78.07	85.77	87.72
2400	74.97	77.47	76.34	77.32	85.34	87.53

Supplementary Table 12. **Effect of sampling rate on intrinsic performance.** Intrinsic evaluation of word2vec (w2v) and fastText word embedding models with various values for the "sampling rate" hyper-parameter. Similarity and relatedness is represented by the Spearman correlation coefficient for the cosine similarity of term pairs and the 3 reference standards. Out-of-vocabulary terms were considered and represented by a null vector for word2vec models. Asterisk (*) indicates default parameter value. Highest performance for each model and each standard is in bold.

Samp	UMNSRS				HDO				XADO			
	w2v		FastT		w2v		FastT		w2v		FastT	
	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel
0	0.561	0.502	0.476	0.472	0.268	0.221	0.295	0.220	0.037	0.103	0.048	0.110
1e-10	0.052	0.138	0.058	0.097	0.023	0.000	0.036	0.003	0.047	0.013	0.048	0.030
1e-9	-0.008	0.003	0.166	0.110	-0.057	-0.024	0.049	0.061	-0.086	0.036	0.020	0.074
1e-8	0.192	0.166	0.401	0.373	0.053	0.076	0.196	0.180	0.037	0.094	-0.031	0.072
1e-7	0.605	0.550	0.556	0.542	0.266	0.192	0.268	0.191	0.013	0.097	-0.007	0.089
1e-6	**0.646**	**0.601**	**0.559**	**0.571**	0.283	0.201	0.302	0.213	**0.055**	0.095	0.049	0.094
1e-5	0.645	0.584	0.544	0.552	**0.288**	0.219	**0.314**	0.216	0.051	0.100	**0.057**	0.103
1e-4	0.609	0.553	0.518	0.509	0.280	0.225	0.309	**0.218**	0.049	0.101	0.056	**0.114**
1e-3*	0.579	0.517	0.502	0.494	0.279	**0.226**	0.306	0.217	0.039	0.102	0.046	0.113
1e-2	0.563	0.507	0.487	0.482	0.272	0.220	0.301	0.216	0.033	**0.105**	0.047	0.113
1e-1	0.558	0.501	0.479	0.479	0.266	0.220	0.295	0.217	0.031	0.102	0.049	0.111

Supplementary Table 13. **Effect of sampling rate on extrinsic performance.** Extrinsic evaluation of word2vec (w2v) and fastText word embedding models with various values for the "sampling rate" hyper-parameter. Accuracies represent F-score for named entity recognition of 3 corpora: BC2GM, JNLPBA, and CHEMDNER. Asterisk (*) indicates default parameter value. Highest performance for each model and each corpus is in bold.

Samp	BC2GM		JNLPBA		CHEMDNER	
	w2v	FastT	w2v	FastT	w2v	FastT
0	**78.10**	78.27	76.48	**78.80**	87.38	89.57
1e-10	69.92	73.09	74.92	75.71	82.44	83.53
1e-9	69.79	76.16	73.84	77.29	81.25	87.24
1e-8	71.54	77.76	76.38	77.37	83.19	88.88
1e-7	76.74	78.04	**78.48**	77.20	86.70	89.10
1e-6	76.64	78.11	76.45	77.77	87.50	89.33
1e-5	77.59	78.57	77.19	78.52	87.84	89.48
1e-4	76.53	78.61	76.97	77.60	87.84	**89.64**
1e-3*	77.73	**80.57**	76.85	78.31	**88.18**	89.45
1e-2	77.66	80.53	76.58	78.07	88.06	89.63
1e-1	77.11	79.76	77.49	77.68	87.78	89.45

Supplementary Table 14. **Effect of learning rate (alpha) on intrinsic performance.** Intrinsic evaluation of word2vec (w2v) and fastText (FastT) word embedding models with various values for the "learning rate (alpha)" hyper-parameter. Similarity and relatedness is represented by the Spearman correlation coefficient for the cosine similarity of term pairs and the 3 reference standards. Out-of-vocabulary terms were considered and represented by a null vector for word2vec models. Asterisk (*) represents default parameter value. Highest performance for each model and each standard is in bold.

Alpha	UMNSRS				HDO				XADO			
	w2v		FastT		w2v		FastT		w2v		FastT	
	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel
0.0125	0.570	0.511	0.466	0.472	0.251	0.204	0.290	0.216	**0.041**	0.106	0.056	0.102
0.025*	0.582	0.523	0.500	0.496	0.272	**0.224**	**0.308**	**0.221**	0.036	0.102	0.049	0.112
0.05	0.601	0.546	0.528	0.517	0.303	0.220	0.306	0.216	0.027	0.110	0.041	**0.118**
0.1	**0.602**	**0.578**	**0.545**	**0.528**	**0.306**	0.204	0.299	0.216	0.025	**0.115**	0.032	**0.118**
0.2	0.008	0.070	0.107	0.091	0.013	0.006	0.048	0.163	0.021	0.035	**0.073**	-0.011
0.5	-0.013	0.062	0.035	0.012	-0.006	0.026	0.047	0.098	0.029	0.015	-0.035	0.028

Supplementary Table 15. **Effect of learning rate (alpha) on extrinsic performance.** Extrinsic evaluation of word2vec (w2v) and fastText (FastT) word embedding models with various values for the "learning rate (alpha)" hyper-parameter. Accuracies represent F-score for named entity recognition of 3 corpora: BC2GM, JNLPBA, and CHEMDNER. Asterisk (*) represents default parameter value. Highest performance for each model and each corpus is in bold.

Alpha	BC2GM		JNLPBA		CHEMDNER	
	w2v	FastT	w2v	FastT	w2v	FastT
0.0125	77.23	78.94	76.79	77.71	87.61	89.55
0.025*	**79.14**	78.85	76.87	**77.80**	87.62	**89.68**
0.05	77.61	78.89	76.87	77.05	**88.04**	88.87
0.1	78.38	**79.02**	**77.45**	77.28	87.85	89.10
0.2	60.53	56.53	68.91	65.27	83.16	76.91
0.5	61.19	51.01	69.17	62.58	82.80	79.42

Supplementary Table 16. Top 5 most similar words to *phosphatidylinositol-4,5-bisphosphate*. fastText identifies syntactically similar terms which also refer to structurally similar molecules. Word2vec identifies less syntactically similar terms but recalls abbreviated forms, where *PIP2* and *PtdIns(4,5)P2* are synonyms to the queried term.

word2vec	fastText
4,5-bisphosphate	phosphatidylinositol-4,5-biphosphate
phosphatidylinositol	phosphatidylinositol-(4,5)-bisphosphate
4,5)-bisphosphate	phosphatidylinositol-4-phosphate
PIP2	4,5-bisphosphate
PtdIns(4,5)P2	phosphatidylinositol-4

Supplementary Table 17. Top 10 most similar words to the rare genetic skeletal malformations disorder *acrodysostosis*. Albright's hereditary osteodystrophy (*Albright_hereditary_osteodystrophy*) and McCune-Albright Syndrome (*McCune_Albright_syndrome*) are genetic disorders effecting the bone, skin and endocrine system. *Dysostosis* and *pycnodysostosis* are disorders concerning bone development.

word2vec	fastText
Albright_hereditary_osteodystrophy	dysostosis
Familial_glucocorticoid_deficiency	pycnodysostosis
McCune_Albright_syndrome	pyknodysostosis
McCune_-Albright	dysostoses
Melnick_Needles_syndrome	spondyloenchondrodysplasia
Hypochondroplasia	hereditary_multiple_exostosis
Hajdu-Cheney	alright_hereditary_osteodystrophy
PHP-la	pseudoachondroplasia
NFNS	chondrodysplasia
PHP1A	macrodystrophia

Supplementary Table 18. Top 5 most similar terms to the out-of-vocabulary genetic variant (SNP): *LRG_1:g.8463G>C*. Reference SNPs are denoted by the *RS-* prefix in their respective database accessions.

rs2243250
rs2241880
rs3212227
rs3212986
rs3748067

Supplementary Table 19. Top 10 words most similar to the SNP *rs2243250*, which is Interleukin 4 – 590C/T polymorphism. In addition to Reference SNP identifiers, word2vec recalls the 590C/T alteration within the first 10 terms, whereas fastText only identifies Reference SNP identifiers.

word2vec	**fastText**
Rs2070874	Rs1800896
Rs1800871	Rs2275913
Rs8193036	Rs2430561
Rs1800872	Rs2241880
Rs20541	Rs2070874
Rs2243248	Rs1800925
Rs2227284	Rs2228145
Rs4711998	Rs1143634
590C>T	Rs1800872
590C/T	Rs1143627

Supplementary Table 20. Top 10 most similar words to *ZNF580* – Zinc Finger Protein 580.

word2vec	**fastText**
hCTGF	ZNF545
Focal_adhesional_kinase	ZNF582
Tmfn2	ZNF521
Deltanp63a	ZNF24
RTEF-1	ZNF202
p-CREB-1	ZNF32
ITGa5	ZNF217
BMP9-dependent	BTG1
Ox-LDL-injured	ZNF281
IGFBP-3-mediated	ZNF703

Supplementary Table 21. Top 10 most similar words to *1,2-dichloroethane* returned by word2vec and fastText models.

word2vec	**fastText**
1,1,1,2,2-tetrachloroethane	dichloroethane
1,1,2-trichloroethane	Water/1,2-dichloroethane
chlorobenzene	1,1-dichloroethane
1,2-dichlorobenzene	1,1,2-trichloroethane
1,4-dioxane	1,2-dichlorobenzene
nitrobezene	chloroethane
dichloroethane	cis-1,2-dichloroethene
Toluene	1,1,2,2-tetrachloroethane
1,1-dichloroethene	trichloroethane
tetrachloroethane	Trans-1,2-dichloroethylene

Supplementary Table 22. Top 10 most similar words to *zinc_finger_protein* returned by word2vec and fastText models.

word2vec	**fastText**
Zinc-finger	Zinc_finger_proteins
ZFP	RET_finger_protein
cGATA-1	Zinc_finger
Neural-restrictive	Ret_finger_protein
Six-zinc	Zinc_fingers
SZF1	Zinc-finger
SP/KLF	ZFP
Six-finger	KRAB
GAGA-like	KRAB-ZFPs
UtroUp	KRAB-ZFP

Supplementary Table 23. Variation (standard deviation) in the intrinsic accuracy for the default parameters. Models were run independently during parameter optimization.

Source Parameter Table	**UMNSRS**				**HDO**				**XADO**			
	w2v		**FastT**		**w2v**		**FastT**		**w2v**		**FastT**	
	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel	Sim	Rel
Window	0.58	0.528	0.497	0.491	0.279	0.224	0.311	0.22	0.032	0.105	0.043	0.113
Negative	0.582	0.524	0.493	0.49	0.277	0.226	0.306	0.218	0.033	0.103	0.046	0.11
Sampling	0.579	0.517	0.502	0.494	0.279	0.226	0.306	0.217	0.039	0.102	0.046	0.113
Min-Count	0.583	0.526	0.496	0.493	0.275	0.229	0.306	0.22	0.039	0.106	0.038	0.11
Alpha	0.582	0.523	0.5	0.496	0.272	0.224	0.308	0.221	0.036	0.102	0.049	0.112
Dimensionality	0.584	0.528	0.498	0.493	0.277	0.223	0.306	0.215	0.034	0.102	0.043	0.113
Epochs	0.58	0.522	0.495	0.495	0.271	0.222	0.307	0.216	0.036	0.106	0.047	0.112
Stdev	0.002	0.004	0.003	0.002	0.003	0.002	0.002	0.002	0.003	0.002	0.004	0.001

Supplementary Table 24. Variation (standard deviation) in the extrinsic accuracy for the default parameters for the word2vec (w2v) and fastText (FastT) models. Models were run independently during parameter optimization.

Source Parameter Table	Word2vec			FastText		
	BC2GM	JNLPBA	CHEMDNER	BC2GM	JNLPBA	CHEMDNER
Window	77.64	77.18	87.74	79.31	77.52	89.38
Negative sub-sampling	78.87	76.67	87.94	78.36	78.23	89.50
Sampling rate	77.73	76.85	88.18	80.57	78.31	89.45
Min-count	77.50	76.8	87.84	79.07	78.59	89.83
Alpha	79.14	76.87	87.62	78.85	77.80	89.68
Dimensionality	77.12	76.83	87.94	79.50	78.10	89.60
Stdev	0.81	0.17	0.19	0.75	0.38	0.16

Supplementary Table 25. Word2vec (w2v) and fastText (FastT) optimized hyper-parameters for each intrinsic standard and extrinsic corpus.

	UMNSRS		HDO		XADO		BC2GM		JNLPBA		CHEMDNER	
	w2v	FastT	w2v	FastT	w2v	FastT	w2v	FastT	w2v	FastT	w2v	FastT
Window	75	75	40	30	1	1	2	2	5	1	1	1
Negative	10	10	15	2	15	10	5	10	15	15	2	10
Sampling	1e-6	1e-6	1e-5	1e-5	1e-6	1e-4	0	1e-3	1e-7	0	1e-3	1e-3
Min-count	0	0	0	5	0	0	5	50	10	5	0	5
Alpha	0.1	0.1	0.05	0.025	0.0125	0.0125	0.025	0.1	0.1	0.025	0.05	0.025
Dim	100	500	100	500	200	400	200	100	200	100	100	100
N_grams	-	5-7 / 6-7	-	6-8	-	3-4	-	2-6 / 3-8	-	3-7	-	3-7

Supplementary Table 26. Intrinsic and extrinsic performance for the word2vec (w2v) and fastText (FastT) embeddings trained with the per-corpus optimized hyper-parameters (Supplementary Table 25). Highest performance for each model and each standard is in bold.

Model	UMNSRS		HDO		XADO		BC2GM	JNLPBA	CHEMDNER
	Sim	Rel	Sim	Rel	Sim	Rel			
w2v	**0.733**	**0.686**	**0.324**	0.234	0.061	**0.110**	76.83	72.03	88.26
FastText	0.541/ 0.581	0.507/ 0.520	0.306	**0.235**	**0.090**	0.089	**79.33** / 78.85	**73.30**	**90.54**

Supplementary Table 27. Optimized word2vec (w2v) and fastText (FastT) hyper-parameters across intrinsic standards and extrinsic corpora.

	Optimized on intrinsic standards		Optimized on extrinsic corpora	
	w2v	FastT	w2v	FastT
Window	30	25	2	1
Negative	15	3	5	10
Sampling	1e-5	1e-5	1e-3	1e-3
Min-count	0	0	10	5
Alpha	0.05	0.05	0.1	0.025
Dim	200	200	50	100
N_grams	-	6-7	-	3-7

Supplementary Table 28. word2vec (w2v) and fastText (FastT) hyper-parameters optimized across intrinsic and extrinsic datasets.

	w2v	FastT
Window	30	25
Negative	15	10
Sampling	1e-5	1e-5
Min-count	0	0
Alpha	0.025	0.025
Dim	200	200
N_grams	-	6-7

Supplementary Table 29. Intrinsic and extrinsic performance for word2vec and fastText embeddings optimized on global optima of hyper-parameters across all corpora and standards (Supplementary Table 28).

	UMNSRS		HDO		XADO		BC2GM	JNLPBA	CHEMDNER
	Sim	Rel	Sim	Rel	Sim	Rel			
word2vec	0.707	0.657	0.296	**0.251**	**0.094**	0.072	*76.04*	*72.74*	87.53
fastText	**0.721**	**0.674**	**0.329**	0.241	0.077	**0.088**	**76.76**	**72.79**	**89.11**