Biomedical information extraction with natural language processing and machine learning methods

Rui Antunes

Electrical Engineering Doctoral Programme

Scientific guidance: Sérgio Matos, Assistant Professor

Department of Electronics, Telecommunications and Informatics Institute of Electronics and Informatics Engineering of Aveiro University of Aveiro

March 7, 2023



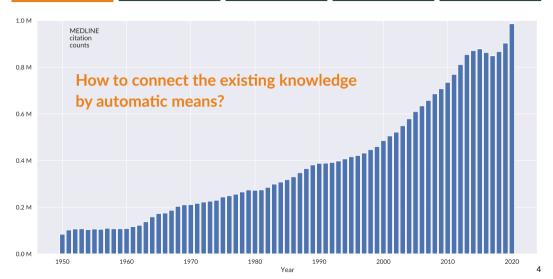


Outline

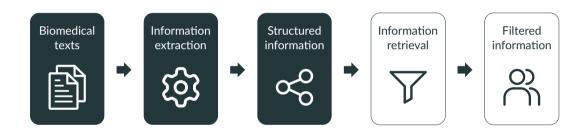
- 1. Introduction
- 2. Biomedical text classification (BioTC)
- 3. Biomedical word sense disambiguation (BioWSD)
- 4. Biomedical relation extraction (BioRE)
- 5. Final remarks

1. Introduction

Introduction: Motivation

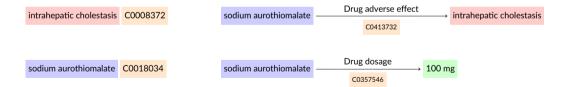


Introduction: Biomedical text mining



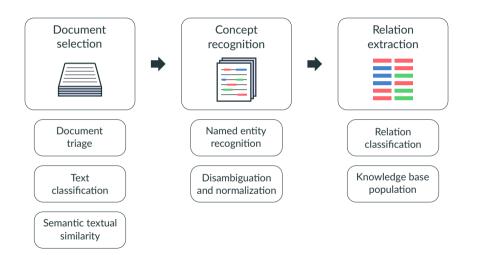
Introduction: Research challenge

A patient with rheumatoid arthritis developed an acute intrahepatic cholestasis after 100 mg of sodium aurothiomalate. Hepatotoxicity in rheumatoid arthritis is well described following a variety of anti-inflammatory preparations but is now a rare complication of gold therapy and the literature on the subject is reviewed.



Source: ADE corpus (PMID 115078)

Introduction: Biomedical information extraction



Introduction: Text encoding

Sparse representation

- · One-hot encoding
- Term frequency-inverse document frequency (TF-IDF)

Dense representation

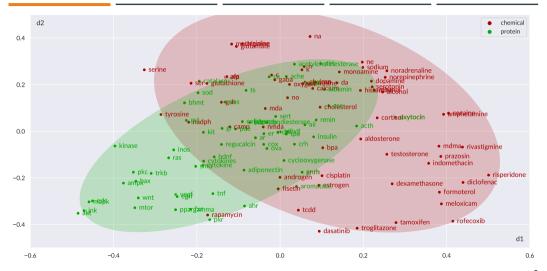
Word embeddings (distributed representations of words)

Mikolov et al. introduced the Skip-gram model, an efficient method for learning high-quality vector representations of words from large amounts of unstructured text data. Unlike most of the previously used neural network architectures for learning word vectors, training of the Skip-gram model does not involve dense matrix multiplications. This makes the training extremely efficient: an optimized single-machine implementation can train on more than 100 billion words in one day.

Source: Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (December 2013).

Distributed representations of words and phrases and their compositionality. NIPS 2013.

Introduction: Word embeddings example



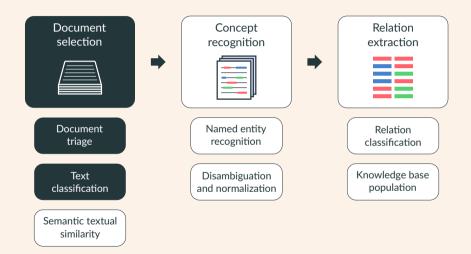
Introduction: Word embeddings example

yellow fever viral disease → ['yellow', 'fever', 'viral', 'disease']

Word	Vector
'yellow'	[-0.163, 0.027,, 0.009]
'fever'	[0.037, -0.169,, 0.062]
'viral'	[-0.037, -0.115,, 0.056]
'disease'	[0.141, -0.107,, -0.031]
Average vector	[-0.006, -0.091,, 0.024]

2. Biomedical text classification (BioTC)

2. Biomedical text classification (BioTC)



BioTC: Literature triage for precision medicine

- BioCreative VI Precision Medicine task dataset
- · Identify articles describing protein-protein interactions affected by genetic mutations

Partition	Abstracts	Positive	Negative
Training	4082	1729	2353
Test	1427	704	723

Source: Doğan, Kim, Chatr-Aryamontri, Wei, Comeau, and Lu (October 2017).

Overview of the BioCreative VI Precision Medicine Track. BioCreative VI Workshop.

https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vi/track-4/

BioTC: BioCreative VI Precision Medicine task

Classical machine learning

Word embeddings (average) · Traditional classifiers (k-NN, LR, MLP)

Deep learning

Word embeddings (concatenation) · Convolutional recurrent neural networks

Syste	em 1	System 2	Syste	em 3
	Embedding layer	• System 1		Embedding layer
	Dropout	 Self-training approach 		Dropout
3x	Convolutional layer	BioCreative III PPI corpus	3x	Convolutional layer
	ReLU activation			ReLU activation
	Average pooling			Average pooling
	Bidirectional LSTM			Dropout
	Dense layer			Bidirectional LSTM
	Sigmoid activation		2x	LSTM
				Dense layer
				Sigmoid activation

BioTC: BioCreative VI Precision Medicine task

	Precision	Recall	F1-score
Classical machine learning			
k-NN (k=99)	0.618	0.553	0.582
LR	0.674	0.546	0.603
MLP	0.606	0.578	0.592
Deep learning			
System 1	0.637	0.681	0.651
System 2	0.640	0.692	0.664
System 3	0.698	0.735	0.715

^{* 5-}fold cross-validation on training set

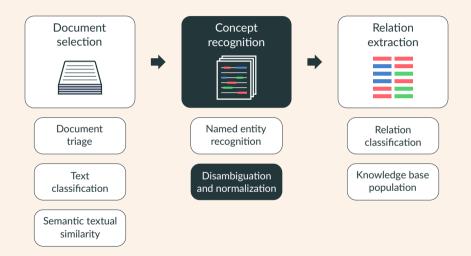
BioTC: BioCreative VI Precision Medicine task

Rank	Work		Average precision	Precision	Recall	F1-score
1	Fergadis, Baziotis, Pappas et al.		0.7158	0.6289	0.7656	0.6906
2	Luo, Yang, Lin, and Wang		0.7253	0.6073	0.7997	0.6904
3	Matos and Antunes	System 2	0.6677	0.5700	0.8736	0.6898
	(ours)	System 1	0.6616	0.5864	0.8338	0.6886
		System 3	0.6929	0.6070	0.7898	0.6864
4	Chen, Singh, Jue, Wang et al.		0.5797	0.5713	0.8253	0.6752
5	Qu, Steppi, Hao, Wang, Lung et al.		0.6632	0.5413	0.8835	0.6713
6	Tran and Kavuluru		0.6439	0.5438	0.8736	0.6703
7	Chen, Panyam, Elangovan et al.		0.6744	0.5361	0.8849	0.6677
8	Altınel, Hüsünbeyi, and Özgür		0.5077	0.5022	0.9801	0.6641
9	Team 405		0.5871	0.5484	0.5710	0.5595
10	Wang, Shen, Elayavilli, Liu et al.		0.4904	0.4649	0.3480	0.3981

* Official test set results

3. Biomedical word sense disambiguation (BioWSD)

3. Biomedical word sense disambiguation (BioWSD)



BioWSD: Problem and dataset

carcinogens in drinking water

Concept	CUI	MeSH
Alcohol drinking	C0001948	D000428
 Water consumption 	C0684271	D004326

MSH WSD dataset

- · 203 ambiguous terms
- · 37 888 ambiguity cases

Source: Antonio Jimeno-Yepes, Bridget T. McInnes, and Alan R. Aronson (June 2011).

Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. BMC Bioinformatics.

https://lhncbc.nlm.nih.gov/ii/areas/MSD/collaboration.html

BioWSD: Approaches

Knowledge-based

Word embeddings · CUI textual definitions · Cosine similarity

Supervised

 $Bag\text{-}of\text{-}words\cdot Word\ embeddings}\cdot Machine\ learning\ classifiers$

BioWSD: Knowledge-based method

carcinogens in drinking water

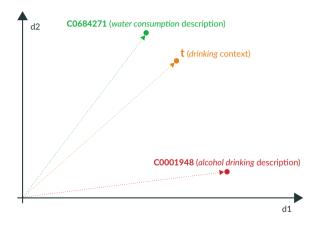
· Ambiguous term: drinking

• Context t: carcinogens in water

Concept CUI

X Alcohol drinking C0001948

✓ Water consumption C0684271



BioWSD: Knowledge-based method

Word embeddings weighting

- Term frequency-inverse document frequency (TF-IDF)
- · Word distance decay (and IDF)

carcinogens in drinking water

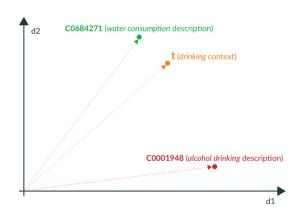
		'carcinogens'	'in'	'water'
	d (distance)	2	1	1
f(d)	1	1.0000	1.0000	1.0000
	1/d	0.5000	1.0000	1.0000
	$\exp\left(-d\right)$	0.1353	0.3679	0.3679
	$1/\ln(1+d)$	0.9102	1.4427	1.4427

BioWSD: Knowledge-based method

Word embeddings · CUI textual definitions · Cosine similarity

Word embeddings averaging functions

score(CUI) = CS(t, CUI)



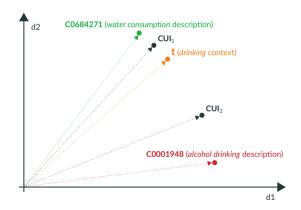
BioWSD: Knowledge-based method extended

Word embeddings · CUI textual definitions · Cosine similarity

Word embeddings averaging functions

CUI-CUI association values (NPMI)

$$\begin{split} & \mathsf{score}(\mathsf{CUI}) = \\ & \frac{1}{N} \sum_{j} \mathsf{NPMI}(\mathsf{CUI}, \mathsf{CUI}_{j}) \cdot \mathsf{CS}(\mathbf{t}, \mathbf{CUI}_{j}) \end{split}$$



BioWSD: Knowledge-based method extended

CUI-CUI association values

- · MeSH (CUI) co-occurrences in MEDLINE citations
- Normalized pointwise mutual information (NPMI)

NPMI	C0000039	C0000052	•••	C4076759
C0000039	1.00	0.23		0.37
C0000052	0.23	1.00		0.68
	•••			
C4076759	0.37	0.68		1.00

BioWSD: Knowledge-based results (accuracy)

CUI-CUI associations	WE Size	Weighting scheme Term frequency	No decay	Fractional	Exponential	Logarithmic
CS	100	0.8321	0.8341	0.8502	0.8278	0.8468
	300	0.8337	0.8365	0.8533	0.8276	0.8501
NPMI ≥ 0.8	100	0.8314	0.8334	0.8493	0.8264	0.8461
	300	0.8332	0.8357	0.8518	0.8264	0.8491
NPMI ≥ 0.5	100	0.8197	0.8236	0.8376	0.8150	0.8343
	300	0.8209	0.8245	0.8396	0.8168	0.8352
NPMI ≥ 0.3	100	0.8600	0.8635	0.8744	0.8459	0.8730
	300	0.8582	0.8611	0.8736	0.8478	0.8719

* 5-fold cross-validation

26

BioWSD: Supervised machine learning results (accuracy)

BoW Features	WE Size	Window	Classifier DT	kNN	LR	MLP	SVM
U	-	-	0.9067	0.9324	0.9205	0.9401	0.9511
В	-	-	0.8335	0.8850	0.8704	0.9224	0.9253
U+B	-	-	0.9019	0.9354	0.9101	0.9445	0.9552
-	100	5	0.9219	0.9452	0.9500	0.9503	0.9449
-		20	0.9185	0.9452	0.9495	0.9498	0.9452
-	300	5	0.9186	0.9449	0.9505	0.9503	0.9452
-		20	0.9186	0.9444	0.9508	0.9514	0.9446
U	100	5	0.9244	0.9464	0.9515	0.9557	0.9490
		20	0.9215	0.9468	0.9514	0.9556	0.9486
	300	5	0.9218	0.9475	0.9519	0.9544	0.9499
		20	0.9194	0.9473	0.9524	0.9550	0.9496

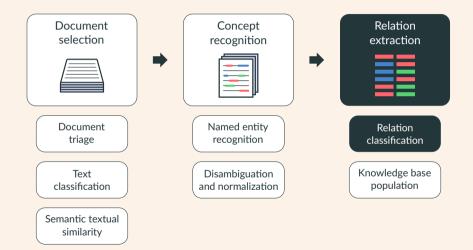
* 5-fold cross-validation 27

BioWSD: Comparison with other works (accuracy)

Work	Approach	Supervised	Knowledge-based
Zhang, Biś, Liu, and He (2019)	Long short-term memory networks	0.9600	-
Pesaranghader, Matwin et al. (2019)	Long short-term memory networks	0.9682	0.9267
Duque, Stevenson et al. (2018)	Co-occurrence graph	-	0.7152
Ours (Antunes and Matos, 2017)	Word embeddings, cosine similarity	0.9557	0.8744
Jimeno-Yepes (2017)	Support vector machine	0.9597	-
Sabbir, Jimeno-Yepes, and Kavuluru (2016)	Word embeddings, k-nearest neighbors	-	0.9434
Tulkens, Šuster, and Daelemans (2016)	Word embeddings, cosine similarity	-	0.84
Jimeno-Yepes and Berlanga (2015)	Word-concept statistical model	0.930	0.891
McInnes and Stevenson (2014)	Semantic similarity measures	0.97	0.78
McInnes and Pedersen (2013)	Semantic similarity measures	-	0.75
Garla and Brandt (2013)	Semantic similarity measures	-	0.8071
Jimeno-Yepes, McInnes, and Aronson (2011)	Naive Bayes classifier	0.9386	0.8383

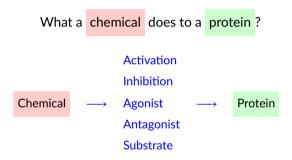
4. Biomedical relation extraction (BioRE)

4. Biomedical relation extraction (BioRE)



BioRE: Text mining chemical-protein interactions

- ChemProt dataset (annotated PubMed abstracts)
- · Relations between chemical compounds (drugs) and genes (proteins)

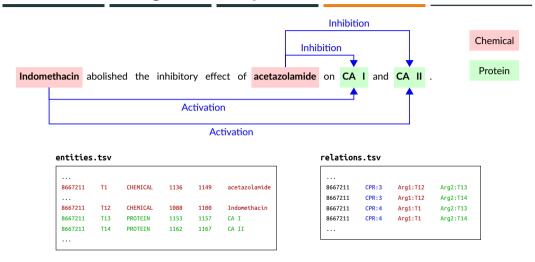


Source: Krallinger, Rabal, Akhondi, Pérez, Santamaría, Rodríguez et al. (October 2017).

Overview of the BioCreative VI chemical-protein interaction track. BioCreative VI Workshop.

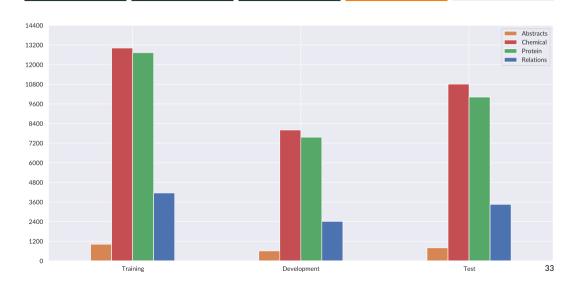
https://biocreative.bioinformatics.udel.edu/events/biocreative-vi/workshop/

BioRE: Text mining chemical-protein interactions

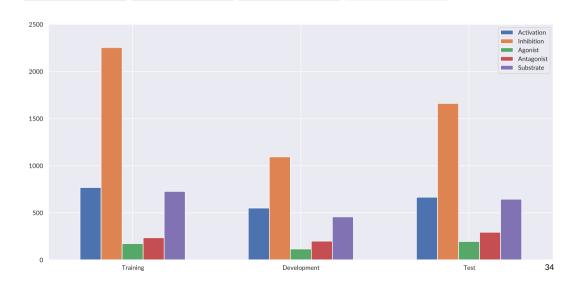


Source: ChemProt dataset (PMID 8667211)

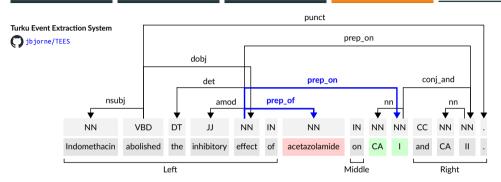
BioRE: ChemProt dataset statistics



BioRE: ChemProt relations statistics



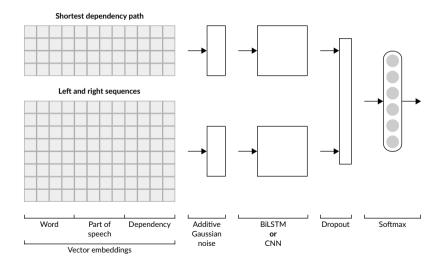
BioRE: ChemProt — Text pre-processing



CPR:4/8667211_7_T1_T13.txt

Shortest dependency path	<pre>chemical NN prep_of effect NN prep_on protein NN </pre>
Chemical left text	Indomethacin NN nsubj abolished VBD the DT det inhibitory JJ amod effect NN dobj of IN
Chemical right text	on IN
Protein left text	on IN
Protein right text	and CC CA NN nn II NN prep_on . . punct

BioRE: ChemProt — Neural network structure



BioRE: ChemProt — Parameters and additional techniques

Gaussian noise standard deviation	0.01
LSTM units	128
LSTM recurrent dropout	0.4
LSTM dropout	0.4
Convolution filters	64
Convolution window sizes	[3, 4, 5]
Dropout rate	0.4
Optimizer	RMSprop
Loss	Categorical cross entropy
Batch size	128
Maximum number of epochs	500
Early stopping patience	30
Early stopping monitor	Validation micro F1-score
Validation split	0.3

 Averaging of three simulations (different random states)

 Balancing between precision and recall (maximize F1-score)

 Additional training data from BioGRID (distant supervision)

BioRE: ChemProt — F1-score on *development* set

(WS, PS, DS)	Features	Model	W	W + P	W + D	W + P + D
(300 ^a , 100, 100)	SDP	BiLSTM CNN	0.6161 0.5642	0.6002 0.5782	0.6473 0.6141	0.6310 0.6092
	LR	BiLSTM CNN	0.5135 0.4293	0.5133 0.4576	0.5209 0.4321	0.5227 0.4216
	SDP + LR	BiLSTM CNN	0.5914 0.5572	0.5812 0.5618	0.6036 0.5672	0.6015 0.5819
(200 ^b , 50, 50)	SDP	BILSTM CNN	0.6229 0.5804	0.6192 0.5841	0.6496 0.6259	0.6453 0.6205
	LR	BiLSTM CNN	0.5030 0.4813	0.5158 0.4504	0.5060 0.4201	0.4849 0.4291
	SDP + LR	BiLSTM CNN	0.5943 0.5690	0.5993 0.5440	0.6126 0.5760	0.5824 0.5605

^a Our PubMed-based word2vec vectors

b BioWordVec fastText vectors by Chen et al. (2019)

BioRE: ChemProt — F1-score on *development* and *test* sets

(WS, PS, DS)			Development	Test
(300 ^a , 100, 100)	Baseline	BiLSTM CNN	0.6473 0.6141	0.6182 0.5932
	Baseline	BiLSTM CNN	0.6496 0.6259	0.6306 0.5959
(200 ^b , 50, 50)	BioGRID	BiLSTM CNN	0.5871 0.5774	0.5964 0.5701
	No validation ^c	BiLSTM CNN	0.6443 0.5547	0.6360 0.5586

^a Our PubMed-based word2vec vectors

^b BioWordVec fastText vectors by Chen et al. (2019)

^c Model trained during 500 epochs (without monitoring)

BioRE: ChemProt — Confusion matrix on *test* set

Predicted	Gold standard						
	Negative	CPR:3 Activation	CPR:4 Inhibition	CPR:5 Agonist	CPR:6 Antagonist	CPR:9 Substrate	Sum
Negative		238	524	97	124	341	1324
Activation	263	382	19	5	0	0	669
Inhibition	401	45	1107	14	2	2	1571
Agonist	45	0	2	79	6	0	132
Antagonist	56	0	1	0	161	0	218
Substrate	185	0	8	0	0	301	494
Sum	950	665	1661	195	293	644	
			True positives	2030			
			False negatives	1428			
			False positives	950			

BioRE: ChemProt — Error analysis

Sentence	Words in the SDP	Predicted	Correct
The introduction of the amino group resulted in not only improved water solubility but also enhanced TLR7 agonistic activity.	group introduction resulted activity	Activation	Agonist
In guinea pigs, antagonist actions of yohimbine at 5-HT(1B) receptors are revealed by blockade of hypothermia evoked by the 5-HT(1B) agonist, GR46,611.	receptors	Agonist	Antagonist
Geldanamycin also disrupts the T-cell receptor-mediated activation of nuclear factor of activated T-cells (NF-AT).	disrupts activation	Activation	Inhibition

BioRE: ChemProt participating teams (F1-score results)

Rank	Authors	Methods	Challenge	Post-challenge
1	Peng, Rios, Kavuluru, and Lu	SVM, CNN and RNN	0.6410	
2	Corbett and Boyle	RNN and CNN	0.6141	0.6258
3	Mehryary, Björne, Salakoski, and Ginter	SVM and RNN	0.6099	0.6310
4	Lim and Kang	Tree-structured RNN	0.5853	0.6410
5	Lung, He, Zhao, Yu, and Zhang	Traditional ML	0.5671	
6	Antunes and Matos	RNN and CNN	0.5181	0.6306
7	Liu, Shen, Elayavilli, Wang et al.	CNN and attention-based RNN	0.4948	0.5270
8	Verga and McCallum	Bi-affine attention network	0.4582	
9	Wang, Yang, Xing, Wu, and Song	RNN	0.3839	
10	Tripodi, Boguslav, Hailu, and Hunter	Traditional ML and neural networks	0.3700	
11	Warikoo, Chang, and Hsu	Tree kernel	0.3092	0.3654
12	Sun		0.2195	
13	Yüksel, Öztürk, Ozkirimli, and Özgür	CNN	0.1864	

5. Final remarks

Final remarks

Key contributions

- · Machine learning models for document classification
- · Knowledge-based methods for entity disambiguation
- · Deep neural networks for relation extraction

Open-source repositories

ruiantunes/2018-n2c2-track-1

Truiantunes/biocreative-vi-track-5-chemprot

Shared tasks

- · 2017 BioCreative VI Track 4 (Precision Medicine)
- 2017 BioCreative VI Track 5 (ChemProt)

- · 2018 n2c2 Track 1 (Cohort Selection for Clinical Trials)
- · 2019 n2c2/OHNLP Track 1 (Clinical Semantic Textual Similarity)
- · 2019 n2c2/UMass Lowell Track 3 (Clinical Concept Normalization)

Final remarks: Tools

Programming	
Python, Bash	
Libraries	
Scikit-learn · Keras (TensorFlow) · XGBoo	$sst \cdot Gensim (word2vec, fastText) \cdot NLTK \cdot NetworkX \cdot TEES$
Models	
BioWordVec · BioSentVec	
Databases	
UMLS · PubMed · MIMIC-III · BioGRID · I	DrugBank - MeSH - RxNorm - SNOMED CT

Final remarks: Thesis publications

Journal articles

- Rui Antunes and Sérgio Matos (December 2017).
 Supervised learning and knowledge-based approaches applied to biomedical word sense disambiguation. Journal of Integrative Bioinformatics, De Gruyter.
- 2. Rui Antunes and Sérgio Matos (October 2019).

 Extraction of chemical-protein interactions from the literature using neural networks and narrow instance representation. Database. Oxford University Press.

Conference articles

- Rui Antunes and Sérgio Matos (October 2016). Machine learning with word embeddings applied to biomedical concept disambiguation. In 22nd Portuguese Conference on Pattern Recognition (Aveiro, Portugal).
- Rui Antunes and Sérgio Matos (June 2017). Biomedical word sense disambiguation with word embeddings. In 11th International Conference on Practical Applications of Computational Biology & Bioinformatics (Porto, Portugal).
- Rui Antunes and Sérgio Matos (October 2017). Evaluation of word embedding vector averaging functions for biomedical word sense disambiguation. In 9th INForum Informatics Symposium (Aveiro, Portugal).
- 4. Sérgio Matos and Rui Antunes (October 2017). Identifying relevant literature for precision medicine using deep neural networks. In BioCreative VI Workshop (Bethesda, Maryland, USA).
- Rui Antunes, João F. Silva, Arnaldo Pereira, and Sérgio Matos (February 2019). Rule-based and machine learning hybrid system for patient cohort selection. In 12th International Joint
 Conference on Biomedical Engineering Systems and Technologies (Prague, Czech Republic).
- Rui Antunes, João F. Silva, and Sérgio Matos (March 2020). Evaluating semantic textual similarity in clinical sentences using deep learning and sentence embeddings. In 35th Annual ACM Symposium on Applied Computing (online).
- João F. Silva, <u>Rui Antunes</u>, João R. Almeida, and Sérgio Matos (April 2020). Clinical concept normalization on medical records using word embeddings and heuristics. In 30th Medical Informatics Europe Conference.

Final remarks: Other publications

Journal articles

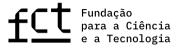
- Sérgio Matos and <u>Rui Antunes</u> (December 2017).
 Protein-protein interaction article classification using a convolutional recurrent neural network with pre-trained word embeddings. Journal of Integrative Bioinformatics, De Gruyter.
- Rezarta Islamaj Dogan, Sun Kim, Andrew Chatr-aryamontri, Chih-Hsuan Wei, Donald C. Comeau, Rui Antunes, Sérgio Matos, et al. (January 2019).
 Overview of the BioCreative VI Precision Medicine Track: mining protein interactions and mutations for precision medicine. Database, Oxford University Press.
- Jorge Miguel Silva, Diogo Pratas, <u>Rui Antunes</u>, Sérgio Matos, and Armando J. Pinho (June 2021).
 Automatic analysis of artistic paintings using information-based measures. Pattern Recognition, Elsevier.
- Tiago Almeida, <u>Rui Antunes</u>. João F. Silva, João R. Almeida, and Sérgio Matos (July 2022).
 Chemical identification and indexing in PubMed full-text articles using deep learning and heuristics. Database, Oxford University Press.
- Robert Leaman, Rezarta Islamaj, Virginia Adams, Mohammed A. Alliheedi, João R. Almeida, <u>Rui Antunes</u>, Robert Bevan, et al. (March 2023).
 Chemical identification and indexing in full-text articles: an overview of the NLM-Chem track at BioCreative VII. Database, Oxford University Press.

Conference articles

- Alina Trifan, Rui Antunes. Sérgio Matos, and José L. Oliveira (April 2020). Understanding depression from psycholinguistic patterns in social media texts. In 42nd European Conference on Information Retrieval (online).
- Alina Trifan, <u>Rui Antunes</u>, and José L. Oliveira (October 2020). Machine learning for depression screening in online communities. In 14th International Conference on Practical Applications of Computational Biology & Bioinformatics (online).
- Rui Antunes, Tiago Almeida, João F. Silva, and Sérgio Matos (November 2021). Chemical-protein relation extraction in PubMed abstracts using BERT and neural networks. In BioCreative VII Challenge Evaluation
 Workshop (online).
- Tiago Almeida, <u>Rui Antunes</u>, João F. Silva, João R. Almeida, and Sérgio Matos (November 2021). Chemical detection and indexing in PubMed full text articles using deep learning and rule-based methods. In BioCreative VII Challenge Evaluation Workshop (online).
- João F. Silva, Tiago Almeida, <u>Rui Antunes</u>, João R. Almeida, and Sérgio Matos (November 2021). Drug mention recognition in Twitter posts using a deep learning approach. In BioCreative VII Challenge Evaluation
 Workshop (online).

Acknowledgments











Work funded by the Portuguese Foundation for Science and Technology (FCT) in the context of the project IF/01694/2013/CP1162/CT0018 and through the doctoral grant SFRH/BD/137000/2018.

Also funded by the Integrated Programme of SR&TD SOCA (Ref. CENTRO-01-0145-FEDER-000010), co-funded by Centro 2020 program, Portugal 2020, Europe Union, through the European Regional Development Fund.





Biomedical information extraction with natural language processing and machine learning methods

Rui Antunes

Electrical Engineering Doctoral Programme

Scientific guidance: Sérgio Matos, Assistant Professor

Department of Electronics, Telecommunications and Informatics Institute of Electronics and Informatics Engineering of Aveiro University of Aveiro

March 7, 2023



