VIRGINIA COMMONWEALTH UNIVERSITY

Nanomedicine Entity Extraction

Bridget McInnes, Ph.D.[1], Ryan Murphy[1], Gabrielle Jones[1], Marley Hodson[2], Tanin Izadi[2], Ivan Jimenez[2], Nastassja Lewinski, Ph.D.[2]

[1] Department of Computer Science, [2]Department of Chemical and Life Science Engineering

Goal

There is a critical need to automatically extract and synthesize knowledge and trends in nanotechnology research from an exponentially increasing body of literature. Engineered nanomaterials (ENMs), such as nanomedicines, are continuously being discovered and Natural Language Processing approaches can semi-automate the cataloging of ENMs and their unique physico-chemical properties; automatically aggregate studies on their exposure and hazards; and link the physicochemical properties to the measured effects.. The goal of this project is to develop a nanomedicine entity extraction system to automatically identify nanomedicine physico-characteristics, exposure and biological effects.

How are nanoparticles described?

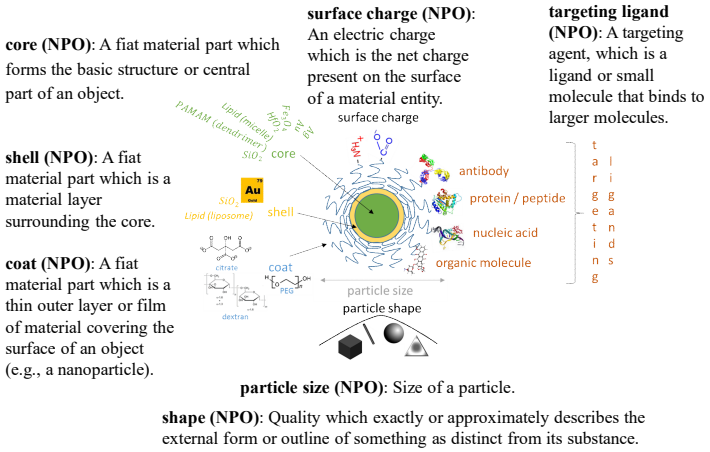
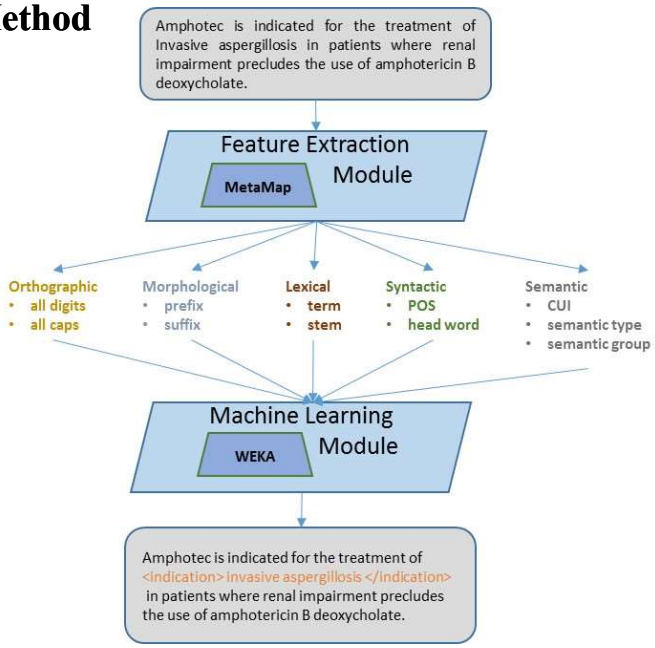


Figure 1. Nanoparticle descriptors with NanoParticle Ontology (NPO) [1] definitions.

Method



END Data

- U.S. Food and Drug Adminstration's Drugs@FDA online database.
- Total number of entities: 28
- Total annotated mentions of entities: ~22,500
- Avg. annotated entities per document: 544 ± 393

Evaluation

- 10-fold cross validation
- Evaluation Metrics: Precision; Recall; F-measure
- Compare with state-of-the-art entity extractions: StanfordNER and OpenNLP

Results

Class	# instances	Entity	Majority Label Baseline	Lexical	Lexical + Syntatic	Lexical + Syntatic + Semantic
Nanoparticle Physico-chemical Characterization	2193	Active Ingredient	0.95	0.89	0.90	
	418	Inactive Ingredient	0.92	0.88	0.90	0.89
	1437	Co-administered Drug	0.95	0.93	0.94	0.90
	904	Nanoparticle	0.95	0.92	0.94	0.93
Exposure	1216	Route of administration	0.96	0.94	0.96	0.95
	2344	Dose	0.91	0.87	0.87	0.86
	923	Frequency	0.91	0.88	0.88	0.91
Biological Response	1549	Indication	0.91	0.86	0.88	0.89
	6885	Adverse Reaction	0.92			

Table 1. F-measure results using Lexical (term), Syntactic (POS) and Semantic (CUI) features into a Naïve Bayes Algorithms

Class	# instances	Entity	Open NLP	StanfordNER
Nanoparticle Physico-chemical Characterization	2193	Active Ingredient	0.62	0.78
	418	Inactive Ingredient	0.12	0.17
	1437	Co-administered Drug	0.46	1.00
	904	Nanoparticle	0.69	0.86
Exposure	1216	Route of administration	0.55	0.79
	2344	Dose	0.34	0.61
	923	Frequency	0.57	0.73
Biological Response	1549	Indication	0.86	0.66
	6885	Adverse Reaction	0.19	0.14

Table 2. F-measure results using the StanfordNER and OpenNLP entity extractors

Conclusion and Future Work

- Current entity extractors are not sufficient to identify all nanomedicine entities.
- Incorporating lexical, syntactic and semantic information aids in the identification of nanomedicine entities.
- Future work is being conducted to evaluate additional features and algorithms.
- The Nanomedicine Entity Extraction system will provide a useful tool to support data mining of nanomedicine literature [2].

References and Acknowledgements

[1] Thomas D, Pappu R, Baker N. "NanoParticle Ontology for cancer nanotechnology research." *J. Biomed. Info.*, 2011, 44, 59-74.
[2] Lewinski N, McInnes B. "Using natural language processing to inform research on nanotechnology." *Beilstein J. Nanotech.*, 2015, 6, 1439-1449.

We thank the VCU Dean's Early Research Initiative program for supporting Gabrielle Jones and the VCU School of Engineering for start up funding to Dr. Lewinski and Dr. McInnes.

We thank Megan Charity for continuing on with this research.

Please contact nalewinski@vcu.edu or btmcinnes@vcu.edu for more information.