V I R G I N I A   C O M M O N W E A L T H   U N I V E R S I T Y

VCU

END, an Annotated Nanomedicine Corpus

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

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

Goal

Annotated corpora are a key resource for Natural Language Processing (NLP) and Information Extraction (IE) methods which employ machine learning. Although annotated corpora are available for pharmaceuticals, resources for nanomedicines are still limited. The goal of this project was to construct a corpus of annotated nanomedicine drug product inserts taken from the U.S. Food and Drug Adminstration's Drugs@FDA online database.

How are nanoparticles described?

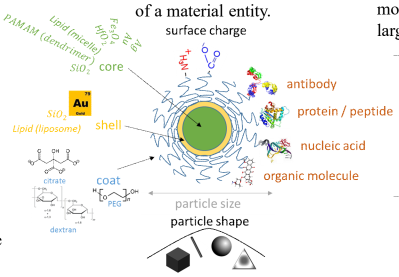
core (NPO): A fiat material part which forms the basic structure or central part of an object.

surface charge (NPO): An electric charge which is the net charge present on the surface of a material entity.

targeting ligand (NPO): A targeting agent, which is a ligand or small molecule that binds to larger molecules.

shell (NPO): A fiat material part which is a material layer surrounding the core.

coat (NPO): A fiat material part which is a thin outer layer or film of material covering the surface of an object (e.g., a nanoparticle).



particle size (NPO): Size of a particle.

shape (NPO): Quality which exactly or approximately describes the external form or outline of something as distinct from its substance.

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

Annotation method

The corpus was annotated by 3 annotators in two stages:
(1) 20 entities were manually identified in each drug product insert, and
(2) identified entities were mapped to controlled vocabularies (ontologies).

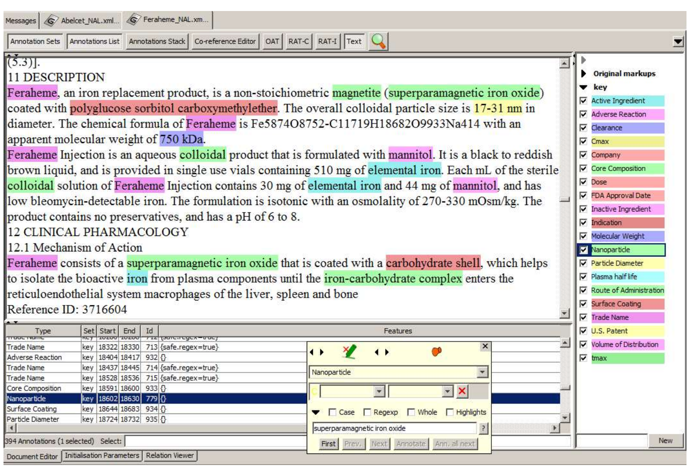


Figure 2. Example GATE annotation.

GATE [2] was chosen as the annotation software because it is open source, widely cited, and used by many NLP researchers.

The annotation guidelines included a list of the entities to be extracted and definitions for the entities (Table 1). The entities were defined as a general term that can be applied to each specific mention contained in the product insert text. Additional guidelines included clarification on how to annotate mentions with multiple meanings, abbreviations, misspelled terms, extra spaces/hyphens/symbols, for annotation consistency.

Results

- Total annotated entities: ~22,500
- Avg. annotated entities per document: 544 ± 393
- 78% of inserts contained nanoparticle specific entities
- 36-74% of inserts contained pharmacokinetic parameters (AUC, Clearance, Cmax, plasma half-life, tmax), often mentioned only once.

Table 1. Select annotated entities with definitions and counts in END corpus.

Entity	Definition	No. mentions
Trade name	Currently used name of the nanomedicine.	6934
Adverse reaction	Non-therapeutic / off target / side effects or toxic injury due to taking the nanomedicine.	4724
Active ingredient	Chemical composition of the agent that is providing the therapeutic effect of the nanomedicine.	2264
Dose	Mass, volume, and/or concentration of nanomedicine administered or any other drug described in the text. Rates are not included (e.g. infusion rate of 5 mg/kg/hr).	2196
Indication	The disease(s)/medical condition(s) for which the nanomedicine is given to detect, treat or prevent.	1502
Route of administration	Method in which the nanomedicine is administered to patients.	1125
Nanoparticle	Type of nanomedicine, which include: Antibody, Antibody-drug conjugate, Dendrimer, Liposome, Micelle, Nanocrystal, Nanoparticle, Polymer, Polymer-aptamer conjugate, Polymer-drug conjugate, Polymer-protein conjugate, Virosome.	716
Surface coating	Chemical composition of the material coating the surface of the nanoparticle core.	53
Molecular weight	Size of the nanomedicine in kilodaltons.	43
Core composition	Chemical composition of the nanoparticle core, which is sometimes also the active ingredient.	42
FDA approval date	Year the nanomedicine was approved for clinical use by the U.S. Food and Drug Administration.	33
U.S. patent	U.S. patent number(s) associated with the nanomedicine.	19
Particle diameter	Size of the nanomedicine in nanometers.	2

Precision, recall and *F-measure* were used as evaluation metrics for entities with the number of mentions > 500.

Table 2. Inter-annotator agreement

	Definition	Equation	Corpus	Nanoparticle
Precision	fraction of correctly labeled entities	$\frac{tp}{tp+fp}$	95%	67%
Recall	fraction of the actual entities that were identified taking into account missing terms.	$\frac{tp}{tp+}$	73%	40%
F-measure	harmonic mean between precision and recall	$2*\frac{Precision*Recall}{Precision+Reca}$	82%	50%

Conclusion

- The clarity of the entity definitions and annotation guidelines (ambiguity resolution) can greatly improve inter-annotator agreement.
- The END corpus could prove to be a valuable resource and promote research in the development of NLP [3] and other machine learning tools to support data mining of nanomedicine literature.

References and Acknowledgements

[1] Thomas D, Pappu R, Baker N. "NanoParticle Ontology for cancer nanotechnology research." *J. Biomed. Info.*, 2011, 44, 59-74.
[2] https://gate.ac.uk/
[3] 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.

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