How are nanoparticles described?

core (NPO): A flat material part which forms the 3D structural or central part of an object.

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

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

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.

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>
<thead>
<tr>
<th>Entity</th>
<th>Definition</th>
<th>Equation</th>
<th>Corpus</th>
<th>Nanoparticle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>fraction of correctly labeled entities</td>
<td>( \frac{tp}{tp + fn} )</td>
<td>95%</td>
<td>67%</td>
</tr>
<tr>
<td>Recall</td>
<td>fraction of actual entities that were identified taking into account missing terms</td>
<td>( \frac{tp}{tp + fp} )</td>
<td>73%</td>
<td>40%</td>
</tr>
<tr>
<td>F-measure</td>
<td>harmonic mean between precision and recall</td>
<td>( \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} )</td>
<td>82%</td>
<td>50%</td>
</tr>
</tbody>
</table>

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