CMAC
FUTURE MANUFACTURING RESEARCH HUB

Automatic Extraction of Pharmaceutical Manufacturing Data from Patents using Natural Language Processing (NLP)

D. Alvarado*, B. Johnston, C. Brown
Future Continuous Manufacturing and Advanced Crystallisation Research Hub (CMAC), University of Strathclyde, Glasgow, UK
*diego.alvarado-maldonado@strath.ac.uk

Introduction

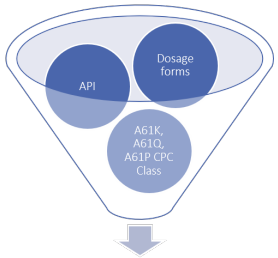
- Deep generative models (DGM) are models capable of generating realistic samples and learning hidden information
- DGM used in drug discovery to generate new molecular entities with desirable biological and chemical properties
- Applications in pharmaceutical manufacturing have not been fully explored
- Potential Benefits of DGM
  - Aid process design by generating a feasible chain of unit operations for the production of an API/dosage forms
  - Improve process understanding through the utilisation of latent variables that may be correlated to process parameters.
- Thousands of data are required to develop a model
- No database that consolidates this information available in literature to be used in DGM for primary or secondary manufacturing domain

Aim

- Propose a methodology to mine data from texts efficiently using natural language processing tools (NLP) and machine learning
- The method focuses on the extraction of **materials, operations and conditions** from **patents**

Methods: Patents Collection

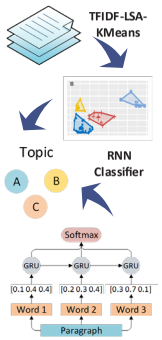
UNITED STATES PATENT AND TRADEMARK OFFICE
uspto
PatentsView USPTO
Learn about patents around the world



> 238K Patents

Methods: Data Extraction Workflow

A. Relevant Paragraph Selection

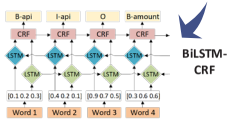


- 49K Patents
- Preprocess text
- Clustering paragraphs (> 3.6 Millions) using machile learning
- Extract keywords
- Group text by topics
- Train DeepNet to classify paragraphs into the identified topics

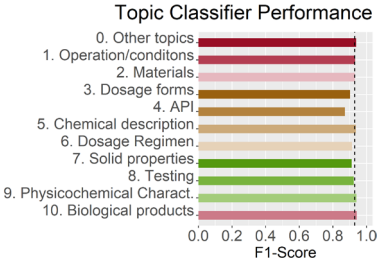
B. Named Entity Recognition (NER)

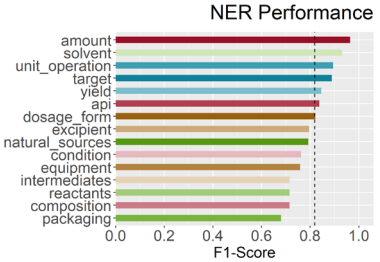
- Select relevant paragraphs (> 5K sentences)
- Annotate words labels (excipient, API, operation, etc)
- Train word vectors using 49K Patents
- Train DeepNet for named entity recognition

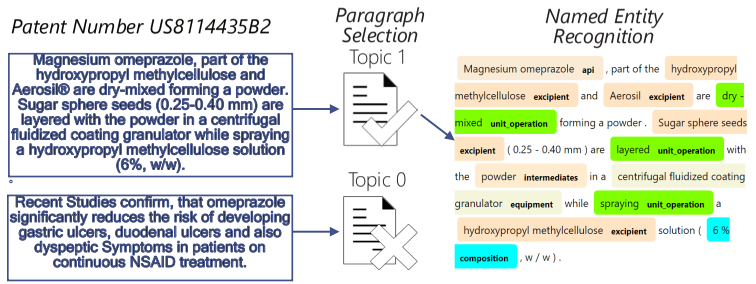
Word	Label
Hydrocodone	B-api
bitartrate	I-api
(	O
48.38	B-amount
g	I-amount
)	O
was	O
stirred	B-operation
in	O
500	B-amount
ml	I-amount
1N	B-composition
NaOH	B-excipient



Results







Discussion

- Information extraction reliability may reach values > 90% and between 60 – 80% for primary and secondary manufacturing, respectively. This result was achieved by applying the proposed method and further post-processing steps.
- Error analysis:
  - Loss of data occurs mainly in materials such as active pharmaceutical ingredients (API), target and intermediates due to low recall
  - Non-entities being recognised as entities, mostly expected for the following categories: intermediates, excipients, composition, unit operations, and dosage forms
  - Other common error is materials classified as another type
- Future work will focus on reducing the errors, improve the proportion of extracted information and structuring a **knowledge graph** to be used in DGM development

References

- Ruthotto, L.; Haber, E. An introduction to deep generative modeling. GAMM Mitteilungen 2021, 44.
- Xia, X.; Hu, J.; Wang, Y.; Zhang, L.; Liu, Z. Graph-based generative models for de Novo drug design. Drug Discov. Today Technol. 2019, 32–33, 45–53.
- Kononova, O.; Huo, H.; He, T.; Rong, Z.; Botari, T.; Sun, W.; Tshitoyan, V.; Ceder, G. Text-mined dataset of inorganic materials synthesis recipes. Sci. Data 2019.

Acknowledgment

EPSRC Centre for Innovative Manufacturing in Continuous Manufacturing and crystallisation