Using *Snomed* to recognize and index chemical and drug mentions

Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, M.Teresa Martín-Valdivia, L. Alfonso Ureña-López
Department of Computer Science, CEATIC, Universidad de Jaén,
{plubeda, mcdiaz, maite, laurena}@ujaen.es

QR

Introduction

- Efficient access to mentions of drugs, medications and chemical entities is a pressing need shared by biomedical researchers, clinicians and pharma industry.
- The recognition of pharmaceutical drugs and chemical entities is a critical step required for the subsequent detection of relations of chemicals with other biomedically relevant entities.

Track 1: NER offset and entity classification

- In the first task, the main objective is to find the chemicals and drugs within the text.

Machine Learning with CRF

- **CRF + bf (basic features)**
  Add features to each word: is lower, is upper, is title, is digit, is alpha, is begin of sentence and is end of sentence.

- **CRF + bf (basic features) + sf (Snomed features)**
  This new feature indicates whether the word is contained in Snomed.

Deep Learning with BiLSTM + CNN

The word embedding used for this task is Spanish Billion Word Corpus.

- BiLSTM + CNN + bf (basic features)

- BiLSTM + CNN + bf (basic features) + sf (Snomed features)
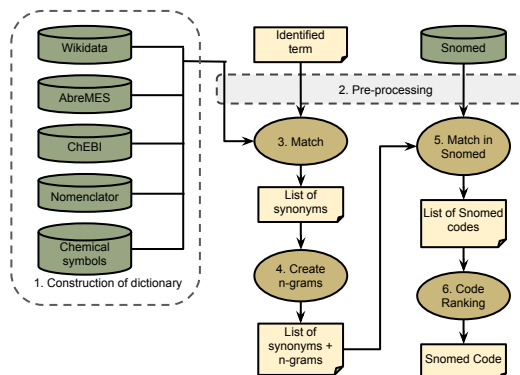
- Results for Track 1:

Run 1	Precision	Recall	F1
CRF + bf	0.926	0.618	0.741
CRF + bf + sf	**0.885**	**0.698**	**0.780**
BiLSTM-CNN + bf	0.844	0.649	0.733
BiLSTM-CNN + bf + sf	0.859	0.696	0.769

Track 2: Concept indexing

- The objective of the second task was to assign a Snomed unique identifier to each concept detected in the previous task.
- System architecture:

- Examples of how the resources and tools applied in the architecture can contribute to the achievement of Snomed concept mapping:

Resource	Input text	Snomed Term	Code
Wikidata	adriamicina	doxorrubicina	372817009
Chemical symbols	Na	sodio	39972003
AbreMES-DB	Hb	hemoglobina	38082009
Hunspell Library	6-Metil-Prednisolona	metilprednisolona	116593003

- Results for Track 2:

Run 1	Precision	Recall	F1
CRF + bf	0.878	0.558	0.682
CRF + bf + sf	**0.852**	**0.633**	**0.736**
BiLSTM-CNN + bf	0.833	0.678	0.683
BiLSTM-CNN + bf + sf	0.828	0.618	0.708

Conclusion

- Adding extra information from Snomed terminology helps classifiers to detect relevant entities within medical texts.
- Apply NLP techniques and tools and the creation of a medical dictionary has contributed to find synonyms for later assigning a single Snomed code.