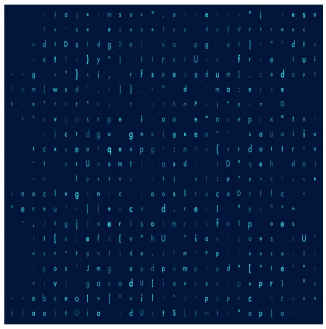
HealthManagement
Promoting Management and Leadership

Enhancing Cross-Lingual Medical Entity Normalisation with xMEN



Medical entity normalisation (MEN) is a fundamental task in clinical natural language processing (NLP), where medical terms identified in free-text documents are linked to standardised medical vocabularies. These vocabularies, such as the Unified Medical Language System (UMLS), serve to ensure semantic consistency across datasets, supporting better data integration and analysis in healthcare research and practice. However, while existing tools like METAMAP and SCISPACY have demonstrated strong results for English-language texts, their performance in other languages often falls short due to limited lexical resources.

The xMEN toolkit aims to address this imbalance by offering a modular system for normalisation of cross-lingual medical entities. Designed to operate effectively across both high-resource and low-resource language scenarios, xMEN combines multilingual alias matching, machine translation and a trainable ranking model to normalise medical terms. By leveraging multilingual representations and machine translation techniques, xMEN can generate training datasets even for languages where annotated resources are scarce.

Modular Design and Methodology
The xMEN toolkit adopts a modular architecture, focusing on two core tasks: candidate generation and candidate ranking. The candidate generation module identifies possible matches for a given medical entity within a target knowledge base. This process employs a hybrid approach, combining a term frequency-inverse document frequency (TF-IDF) model with SAPBERT, a multilingual biomedical language model trained on UMLS data. TF-IDF prioritises surface-level string matching, while SAPBERT enhances semantic understanding through pre-trained embeddings.

Following candidate generation, the candidate ranking module refines the list of potential matches using a transformer-based cross-encoder model. This model, capable of being trained specifically for the task, evaluates both the identified medical entity and its context, helping to distinguish between ambiguous terms. For example, the term "Paget's disease" could refer to multiple conditions, including Paget's disease of the bone or Paget's disease of the breast, requiring contextual understanding for correct disambiguation.

A unique feature of xMEN is its ability to handle low-resource language scenarios. In contexts where labelled datasets are limited, the toolkit uses machine translation to create weakly supervised datasets. English-language datasets are translated into the target language, and entity annotations are projected onto the translated text, enabling the creation of a weakly labelled training set. This allows the cross-encoder model to be trained even in languages lacking sufficient gold-standard data.

To further refine the candidate ranking, xMEN introduces a novel rank regularisation technique. This approach adjusts the balance between general-purpose candidate generation and task-specific re-ranking during model training. By incorporating both semantic similarity and task-specific patterns, rank regularisation enhances performance without overfitting a specific dataset.

Performance Across Multiple Languages
xMEN's effectiveness has been validated across several benchmark datasets, including MANTRA GSC, QUAERO and BRONCO150, covering languages such as English, Spanish, French, German and Dutch. The toolkit consistently achieved state-of-the-art results across these datasets, surpassing existing methods like METAMAP and SCISPACY in terms of accuracy and recall.

A key strength of xMEN is its ability to perform well in both high-resource and low-resource scenarios. In high-resource languages, such as English and Spanish, the cross-encoder achieved significant performance improvements when trained on large annotated datasets. In lower-resource languages, xMEN leveraged machine translation to create effective weakly supervised datasets, achieving notable accuracy gains

compared to baseline models.

However, challenges remain when dealing with complex entity mentions. Longer or multi-token terms, such as "malignant neoplasm of the stomach," often resulted in reduced recall during candidate generation. Lexical ambiguity, where multiple concepts share the same alias, also presented difficulties. Despite these challenges, xMEN's modular architecture allowed for continuous improvement, with its rank regularisation technique helping mitigate some of these issues.

Open-Source Accessibility and Real-world Applications
xMEN has been released as an open-source Python toolkit, ensuring accessibility for researchers and developers in the global medical community. The toolkit provides both a Python API and a command-line interface, making it adaptable to a wide range of data pipelines and workflows. Researchers can use it for tasks such as medical terminology mapping, clinical data standardisation and multilingual health data analysis.

A significant advantage of xMEN is its compatibility with standardised data formats, such as BIGBIO, a schema used for biomedical NLP benchmarks. This ensures interoperability with various datasets and facilitates the comparison of results across different studies. Additionally, the toolkit includes pre-processing and evaluation tools, along with clearly defined configurations for each task, ensuring reproducible benchmarking.

Beyond research, xMEN holds substantial potential in real-world applications. Healthcare organisations working with multilingual patient records can benefit from its ability to standardise medical terms across languages, improving data interoperability and facilitating international research collaborations. The toolkit's capacity to function in both high- and low-resource scenarios makes it particularly valuable for regions where data resources are limited, yet accurate medical data standardisation is crucial.

The xMEN toolkit represents a significant advancement in the field of cross-lingual medical entity normalisation, addressing the challenges of language barriers in clinical text processing. Its modular architecture, combination of TF-IDF and SAPBERT for candidate generation and a trainable cross-encoder with rank regularisation ensure strong performance across both high-resource and low-resource languages.

By integrating machine translation to create weakly supervised datasets and maintaining compatibility with standardised formats, xMEN provides a scalable and adaptable solution for the global medical research community. Its open-source availability encourages collaboration and innovation, while its demonstrated performance across multiple languages highlights its robustness.

While challenges such as complex entity mentions and lexical ambiguity remain, xMEN's flexible design allows for continuous improvement. Future developments could focus on expanding language coverage, refining ranking techniques and addressing entity complexity. Overall, xMEN's contributions offer a promising path towards more inclusive and effective medical data standardisation worldwide.

Source: JAMIA Open

Image Credit: iStock

Published on : Fri, 10 Jan 2025