Specialised Methods in NLP and AI

BGSynthMedic: Synthetic Parallel English-Bulgarian Discharge Summary Corpus with Automatic Annotations

Svetla Boytcheva, Sylvia Vassileva, Pavel Boytchev
Faculty of Mathematics and Informatics
Sofia University St. Kliment Ohridski
Sofia, Bulgaria
svetla@uni-sofia.bg, svasileva@fmi.uni-sofia.bg

Abstract

Data scarcity is a significant problem for advancing clinical Natural Language Processing (NLP) in low-resource languages like Bulgarian. To address this issue, we present BGSynthMedic, a publicly available synthetic parallel corpus of English–Bulgarian discharge summaries with automatic annotations for five entity types and their Unified Medical Language System (UMLS) concept links. We use the SynthMedic dataset, a corpus of 450 synthetic discharge summaries generated using medical guidelines and GPT-4, and automatically translate it into Bulgarian with BgGPT. We use zero-shot multilingual and monolingual Named Entity Recognition (NER) models from the GLiNER family to automatically label diseases, symptoms, procedures, medication, and observations in the original and translated summaries. The extracted entities are automatically linked to the UMLS. The machine translation quality is evaluated using reference-free metrics, yielding a BERTScore of 0.90 and a LaBSE of 0.89. The presented pipeline demonstrates how to create a parallel corpus of synthetic discharge summaries and a set of silver-standard annotations in a low-resource language. We release the BGSynthMedic dataset, including parallel medical term dictionaries and abbreviations, as a community resource to facilitate future research in Bulgarian clinical NLP.

Keywords: BioNLP, Silver corpus, LLM, Synthetic corpus, Machine Translation, LLM, NER, Entity Linking, UMLS

1 Introduction

Millions of clinical documents are created in the healthcare system in unstructured text format. These documents contain personal and sensitive information, and there are many limitations to the automatic processing of such documents. Due to these limitations, advancement of biomedical and clinical Natural Language Processing (NLP), especially for low-resource languages, is slower than in other domains (Névéol et al., 2018). Clinical Named Entity Recognition (NER) is crucial for extracting structured information from clinical documents, thereby significantly improving process automation in healthcare and positively impacting patient care. The complexity of clinical language and terminology, combined with multilingual documentation and abbreviations, poses significant challenges. There is increasing demand for the development of annotated clinical text corpora, and very limited availability of such annotated datasets.

Large Language Models (LLMs) have transformed multilingual NER by moving the focus from classification tasks to generation and effective cross-lingual transfer learning. However, transformer models still outperform general and domain-adapted LLMs on the clinical NER task (Lu et al., 2025; Monajatipoor et al., 2024) because LLMs struggle to detect multi-word entities and precisely determine span boundaries for complex medical terminology. In the general-domain NER setting, LLMs achieve competitive F1-scores compared to BERT-family transformers, but in clinical texts, the opposite trend is observed.

The current state-of-the-art multilingual models for clinical NER that cover the major European languages are mDeBERTaV3 (He et al., 2021), XLM-RoBERTa (Conneau et al., 2019), Medical mT5 (García-Ferrero et al., 2024), GLiNER-multi-v2.1 (Zaratiana et al., 2023), and GLiNER-biomed-large-v1.0 (Stepanov and Shtopko, 2024). These models demonstrate F1-score across various benchmark datasets for clinical NER in English, reaching a range of 0.85-0.89, while for other languages like Italian, Spanish, and French, the performance drops to 0.62-0.75 (García-Ferrero et al., 2024). The ecosystem of BioASQ / CLEF / BioCreative challenges plays a key role in benchmarking for

multilingual clinical NER. The evaluation results in the challenges DisTEMIST (disease) (Miranda-Escalada et al., 2022), SympTEMIST (symptoms, signs, findings) (Lima-López et al., 2023), MultiCardioNER (diseases and drugs) (Lima-López et al., 2024), and MedProcNER (procedures) (Lima-López et al., 2023) show that monolingual BERT-family models demonstrate high accuracy for these tasks. In SympTEMIST, the best result, 0.7477, for the NER task was achieved by an ensemble of fine-tuned pretrained models for BSC-Bio-Es, Roberta-Biomedical-Es, XLMR-Galen, BETO, and mBERT-Galen. For MedProcNER, the best F1 Score, 0.7985, was achieved by a transformer-based model with masked CRF and data augmentation. In DisTEMIST, the highest F1 score, 0.777, was achieved by the fine-tuned PlanTL-GOB-ES/roberta-base-biomedical-clinical-es model. In MultiCardioNER, the best F1 Score of 0.8199 was also achieved using an ensemble of RoBERTa-based models.

Several silver-annotated corpora were developed in the biomedical domain (Sousa et al., 2019). LLMs are used not only for the NER task, but also for machine translation of clinical texts (Anyaegbuna et al., 2026).

This motivated us to propose an approach to machine translation, NER, and Entity Linking for clinical text corpora in low-resource languages.

In this paper, we use the SynthMedic corpus (Grazhdanski et al., 2025) as the source corpus of synthetically generated clinical texts. The targeted language for machine translation of the corpus is Bulgarian. For annotation, five categories are used: disease, symptom, procedure, medication, and observation. In the proposed methodology (see Fig. 1), we first employ a few-shot prompting technique with a large language model to translate synthetic discharge summaries from English into Bulgarian, generating high-quality bilingual medical texts. A human expert checks a small subset of translations to validate their quality. The translated documents are subsequently annotated using multiple zero-shot pretrained monolingual and multilingual NER models to identify relevant medical entities. Additionally, the original English source documents undergo the same annotation process. For further validation, we compare entity annotations across the English and the Bulgarian corpora and ground the extracted entities in a controlled medical vocabulary via entity linking, enabling rigorous assessment of the consistency and accuracy of assigned entity categories across languages. We provide the parallel corpus as a community resource.[1]

2 Data

For this study, we use the SynthMedic corpus (Grazhdanski et al., 2025), which contains 900 synthetic English discharge summaries across five medical specialties and nine socially significant diseases. Each disease has 100 LLM-generated discharge summaries - 50 by GPT-4 and 50 by Llama 3.1 8b. Instruct. The dataset contains a wide variety of diseases, symptoms and signs, treatments, and tests, including about 4K diseases, 15K symptoms, 14K treatments, and 13K tests.

We briefly summarize the evaluation of the dataset performed by Grazhdanski et al. (Grazhdanski et al., 2025). The dataset was evaluated using LLM approaches, including a faithfulness metric that checks consistency with the underlying medical references, as well as knowledge-graph–based validation and automatic correction. The faithfulness of the whole corpus was reported as 93.65%; however, the GPT-4-generated summaries were scored about 10% higher than those generated by Llama. In addition, human experts reviewed a subset of the dataset and evaluated it using a rubric, yielding a System Usability Score of 94.35%. Human experts judged the synthetic summaries to be credible and factually accurate when the generation was grounded in Merck Manual[2] references.

SynthMedic is a publicly available dataset covering multiple diseases, released under a permissive license, making it well-suited for creating a parallel corpus in a lower-resource language such as Bulgarian. Furthermore, since it was recently released, existing LLMs may not yet have been trained on it. For this study, we use GPT-4-generated discharge summaries because they have been reported to be of higher quality.

3 Methods

3.1 Translation Pipeline

The SynthMedic corpus was translated into Bulgarian using LLMs. To refine the prompt and select the best model, we conducted initial experiments using a small excerpt (about 5%) of the SynthMedic corpus, including several discharge summaries for each disease. This allowed us to investigate the models'

[1] https://github.com/svassileva/bgsynthmedic
[2] https://www.merckmanuals.com/professional

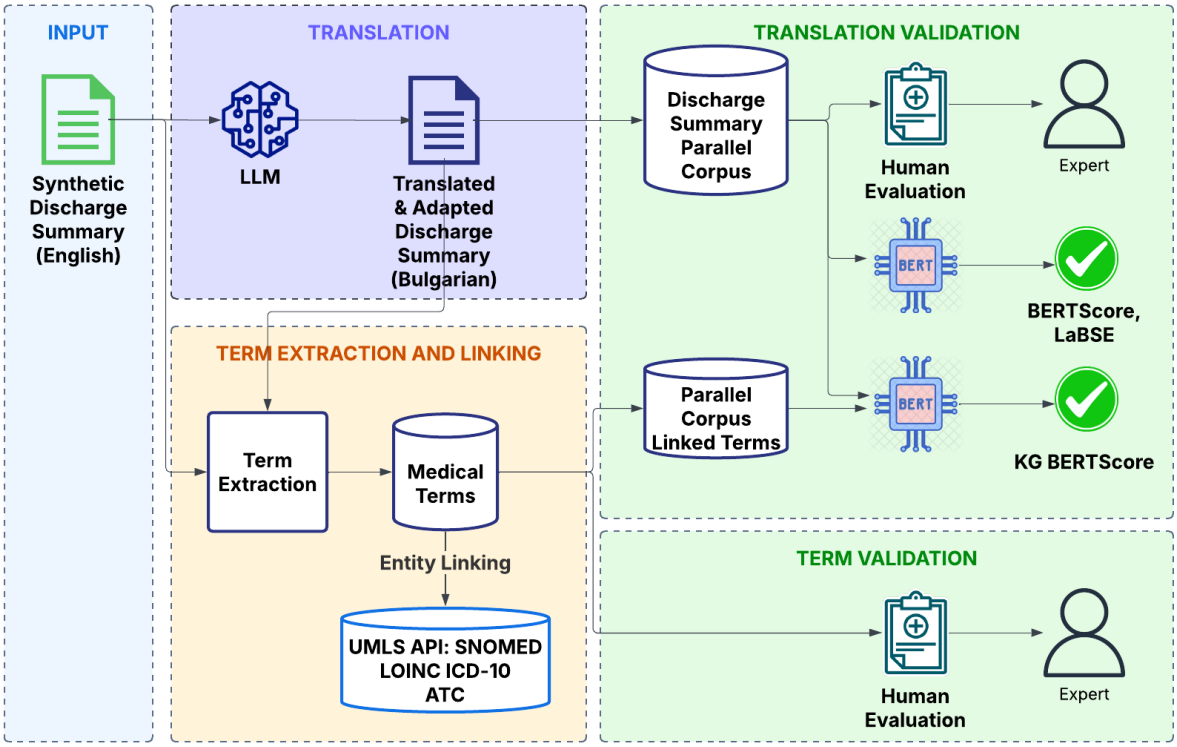

Figure 1: The parallel corpus generation and validation process.

performance in translating specific medical terminology for each clinical specialty. Each document was translated using the same initial prompt using GPT-4.1 (temperature=0.7), GPT-4.1-mini, GPT-4.1-nano[3] (using the API), and Gemma3[4] (using ollama, temperature=0.1, num_predict=7000). The prompt also required generating a parallel dictionary of all medical terms from the source document and the generated translation. The prompt also requested an abbreviation dictionary that includes abbreviations in the source document, their expansion, the proposed target abbreviation (if any exists in the target language), and its expansion in the target language. A human evaluated all translations from the initial experiments and the generated parallel dictionaries. The results identified major issues with the translation of medical terms and abbreviations, as well as with the style of the generated discharge summaries. Further experiments with prompt and parameter modifications did not yield significant improvements in the quality of medical term translation, even with few-shot prompts. To address this issue, we switched to a specific model trained on the Bulgarian general domain: BgGPT (bggpt-gemma-3-27b-fp8)[5]. After selecting the model and adjusting the prompt based on manual testing, we arrived at the final version used to generate all translations. The final few-shot prompt (see Appendix D) includes examples of translations for the most common and problematic medical terms and abbreviations to guide consistent translations of terminology. It also instructs the model to adjust the style of the discharge summary translation to be more formal and to address the patient's general practitioner rather than the patient, which underscores a difference in medical documentation in Bulgaria compared with English-speaking countries. Each discharge summary is translated using a single model call using the BG-GPT API. The final version of the translations was generated with the specified prompt and parameters max_tokens=16384 and temperature=0.7.

3.2 Sentence Alignment and Corpus Statistics

Table 1 presents statistics for the original English dataset and its automatic translation. Sentence splitting was performed using heuristics, including ignoring a small list of abbreviations like "Dr.", or *"г."* (Bulgarian abbreviation for year in a date). Sentence

[3] https://chat.openai.com
[4] https://ollama.com/library/gemma3
[5] https://huggingface.co/INSAIT-Institute/BgGPT-Gemma-3-27B-IT

alignment was performed using the Gale-Church algorithm (Gale and Church, 1993) implemented in the nltk library.[6] The algorithm uses only sentence lengths in characters and a simple probabilistic model to find an optimal alignment between sentences in a parallel corpus. The alignment was performed at the sentence level, thus creating a parallel sentence corpus.

Metric	EN	BG
Documents	450	450
Total sentences	22,252	22,252
Total words	253,675	265,980
Avg words per sentence	11.40	11.95
Avg sentences per document	50.57	50.57
Avg words per document	576.53	604.50

Table 1: Statistics for the original and translated datasets.

3.3 Corpus Annotation

We used zero-shot methods to automatically annotate the parallel corpus with entities and their UMLS codes in English and Bulgarian discharge summaries.

3.3.1 Named Entity Recognition

We experimented with several zero-shot NER models from the GLiNER family: Biomedical GLiNER (Yazdani et al., 2025), GLiNER-X-Large[7], GLiNER Large v2.1 (Zaratiana et al., 2023), and GLiNER multilingual MoE. The models' details are shown in Table 2. For all experiments, we used the following entity types: "disease or medical condition", "abnormality", "symptom or sign", "medical procedure or clinical test", "treatment", "medication", "clinical observation or diagnostic finding". As a post-processing step, these labels were mapped to the five main entity categories: "disease", "procedure", "symptom", "observation", and "medication". The precise definition of labels allowed us to disambiguate some medical terms that, depending on the context, can be classified into different categories (see Appendix C). The GLiNER model allows selecting a custom threshold, which determines whether the span represents an entity. Based on initial experiments, we selected the entity threshold to be 0.2 to identify as many entities as possible. We applied the same models to the English source texts and the Bulgarian translations using the same thresholds to produce a silver-standard annotation corpus. We release the annotations generated by all four models separately. Since GLiNER models use span-based classification for identifying entities, the annotations support nested or overlapping entities of different types.

Model	Multilingual	Domain	BG support
Biomedical GLiNER	No	Biomedical	No
GLiNER-X-Large	Yes	General	No
GLiNER Large v2.1	No	General	No
GLiNER multilingual MoE	Yes	General	Yes

Table 2: Overview of different NER models used in the experiments.

For evaluation, we used a zero-shot NER model - GLiNER multilingual MoE[8]. The model supports multiple languages, including Bulgarian, and is based on GLiNER, which has shown strong performance as a zero-shot NER model across different domains. We used the categories as labels – disease, symptom, medication, procedure, and observation. The threshold for entity detection was set to 0.5.

3.3.2 Entity Linking

UMLS is one of the widely used controlled vocabularies in clinical medicine (Bodenreider, 2004). It comprises more than 1 million biomedical concepts and 5 million concept names. It covers over 100 controlled medical vocabularies, biomedical ontologies, and standard classifications. We performed automatic entity linking to UMLS for the extracted entities in both languages.

We used a cross-lingual SapBERT (cambridgeltlSapBERT-UMLS-2020AB-all-lang) (Liu et al., 2021) model as a baseline model for entity linking, employing cosine similarity search. SapBERT is the standard model for cross-lingual entity linking with the Unified Medical Language System (UMLS)[9] and has shown strong zero-shot performance and good generalization across many languages (Liu et al., 2021). To link English entities, we generated their SapBERT embeddings and performed a cosine similarity search in a knowledge base of UMLS terms, filtered by entity category.

[6] https://www.nltk.org/api/nltk.translate.gale_church.html

[7] https://huggingface.co/knowledgator/gliner-x-large

[8] https://huggingface.co/Mayank6255/GLiNER-MoE-MultiLingual

[9] https://www.nlm.nih.gov/research/umls/index.html

For each entity category, we defined a mapping to UMLS semantic types to reduce the candidate-entity search space (see Appendix A). Based on the filtered concepts, we compiled all their synonyms from the UMLS knowledge base (version 2025AB). We selected the top entity from the knowledge base based on cosine similarity using a FAISS index[10] and assigned the entity code to its mention in the text if the similarity exceeded 0.85. To link entities in Bulgarian, we used automatic translation with the BgGPT LLM[11] before linking them to UMLS, since UMLS is not available in Bulgarian. The entity translation prompt is shown in Appendix B. It translates entity spans in batches for efficiency without any additional context. The model version was bggpt-gemma-3-27b-fp8, and the temperature was set to 0 to reduce the risk of hallucinations and to facilitate reproducibility. The SapBERT cosine similarity search provides a baseline for entity linking and is also used in translation evaluation.

The UMLS API[12] was used as an additional method for automatically validating the extracted terms from NER and their associated categories. We searched for each extracted entity from the NER model using the UMLS API. If the entity was an "exact match", the extracted term was guaranteed to be a medical term from a standard medical vocabulary or ontology. We applied additional validation by comparing the associated entity category with the semantic type of the mapped UMLS term. The UMLS semantic types are narrower than the five broader categories in the scope of our research, so we used a predefined mapping between the two sets of categories (see Appendix C). We also performed an entity search in UMLS using additional parameters such as *words, normalizedString, normalizedWords, rightTruncation, and leftTruncation* for partial matching with lower confidence to cover scenarios such as plural forms of terms, other linguistic variants, and paraphrases. This also covers specific expressions classified under broader UMLS terms. For instance, "RT-PCR test" is mapped in UMLS to "SARS-CoV-2 RT-PCR test"[13], which can be missed if the search was limited to the *exact* match search. We generated the silver-standard entity-linking annotations using the UMLS API, as it is a standard method for this task.

4 Evaluation

4.1 Corpus Evaluation

Creating a parallel reference corpus for evaluating machine translation requires significant effort from bilingual human experts. In the absence of a gold-standard corpus for evaluation, we used reference-free metrics for quality estimation, such as BERTScore (Zhang et al., 2019) and LaBSE (Feng et al., 2022). They measure the semantic similarity between the source and hypothesis translation in a shared embedding space. We used XLM-RoBERTa (xlm-roberta-base) to encode the English and Bulgarian sentences in a shared semantic space and compute token-level cosine similarities to create a cross-lingual alignment. Finally, BERTScore precision, recall, and F1 metrics were calculated to assess how well the text's meaning was preserved in the Bulgarian translation. We used the BERT-score library[14] implementation to calculate the BERTScore metric.

BERTScore Definition. Let the reference sentence be $x = \langle x_1, x_2, \ldots, x_k \rangle$ and the candidate sentence be $\hat{x} = \langle \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m \rangle$. Contextual embeddings are obtained for each token $\mathbf{x}_i \in \mathbb{R}^d$, and for each token $\hat{\mathbf{x}}_j \in \mathbb{R}^d$, from a pretrained language model. Token similarity is calculated using cosine similarity $\cos(\mathbf{x}_i, \hat{\mathbf{x}}_j)$. The BERTScore-Precision, Recall, and F1 Score are calculated using Equations 1, 2, 3 (Zhang et al., 2019).

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \cos\big(\mathbf{x}_i, \hat{\mathbf{x}}_j\big) \qquad (1)$$

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \cos\big(\mathbf{x}_i, \hat{\mathbf{x}}_j\big) \qquad (2)$$

$$F_{\text{BERT}} = \frac{2\, P_{\text{BERT}}\, R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \qquad (3)$$

LaBSE (Language-agnostic BERT Sentence Embedding) (Feng et al., 2022) is a publicly available model trained to cluster semantically equivalent sentences across different languages in a shared embedding space. The cosine similarity between the source and the hypothesis sentence pairs is treated as a sentence-level quality score. High cosine similarity between sentences in different languages

[10] https://github.com/facebookresearch/faiss

[11] https://huggingface.co/INSAIT-Institute/BgGPT-Gemma-3-27B-IT

[12] https://documentation.uts.nlm.nih.gov/rest/home.html

[13] https://uts.nlm.nih.gov/uts/umls/concept/C5400530

[14] https://pypi.org/project/bert-score/

should indicate that the translation preserves the original meaning. LaBSE similarity has been used as a strong baseline in recent works for automatic MT evaluation (Mujadia et al., 2023). We use the sentence-transformers/LaBSE model for calculating the metric.

KG-BERTScore (Wu et al., 2022) is an extension of the BERTScore metric that combines standard BERTScore with bilingual named-entity matching using a knowledge graph. It has shown a higher correlation with human judgments. KG-BERTScore aims to ensure that the same important entities are preserved in the translation. We applied KG-BERTScore by automatically extracting five named entity types from the discharge summaries (disease, symptom, medication, procedure, and observation) and aligning the translated entities in two ways: either by linking them to the UMLS (Bodenreider, 2004) or by using semantic alignment. We used the methods described in Section 3.3 to detect and link entities.

KG-Score Definition. Let the source sentence be $s = \langle s_1, s_2, \ldots, s_n \rangle$ and the machine-translated sentence be $t = \langle t_1, t_2, \ldots, t_m \rangle$. We extract named entities from both sentences using a zero-shot NER model, and denote the sets of entities by $\text{entities}(s)$ and $\text{entities}(t)$, respectively. For each entity in $\text{entities}(s)$, we try to find a corresponding entity in $\text{entities}(t)$ by matching their UMLS codes or SapBERT entity embeddings, resulting in a set of successfully matched entities $\text{matches}(\text{entities}(s), \text{entities}(t))$. The KG-Score F_{KG} is defined as the ratio between the number of matched entities and the number of source entities, as shown in Equation 4. For KG-Score-UMLS, we used the linked entities in each language to identify corresponding entities across languages that map to the same UMLS concept. This allowed us to calculate entity-level precision, recall, and F1 score, and to combine them with the original BERTScore. For KG-Score-Semantic, we used semantic alignment via cosine similarity search with SapBERT embeddings to match the closest entity in the translation, regardless of entity type. We set a threshold of 0.85 to match entities and greedily align entity pairs between English and Bulgarian.

$$F_{\text{KG}} = \frac{\text{matches}\big(\text{entities}(s), \text{entities}(t)\big)}{\text{entities}(s)} \quad (4)$$

KG-BERTScore Definition. KG-BERTScore linearly combines the KG-Score F_{KG} with the BERTScore F_{BERT}. Given a combination weight $\alpha \in [0, 1]$, the KG-BERTScore is defined as $F_{\text{KG-BERT}} = \alpha F_{\text{KG}} + (1 - \alpha) F_{\text{BERT}}$. We set the α value to 0.5, as in the original paper (Wu et al., 2022).

Using the two different KG-Score strategies, we report KG-BERTScore-UMLS and KG-BERTScore-Semantic.

4.2 Annotation Evaluation

We evaluated the quality of the NER annotations by comparing them with a small subset of documents manually labeled by a human expert across the five target categories. One annotator labeled each document, producing labels for 27 documents per language (3 documents per diagnosis per language). Table 3 shows the number of entities based on entity type in the human-labeled documents. We publish the annotation guidelines together with the dataset and labeled documents. The annotators used the NER Annotator for SpaCy tool[15]. We evaluate the strict precision, recall, and F1-score of all the GLiNER models on the manually labeled dataset. We also compared the number of extracted entities per entity type by the Biomedical GLiNER model from the source and translated texts. In addition, we measured the number and percentage of entities successfully linked using the UMLS API.

Entity type	Number
DISEASE	298
MEDICATION	186
OBSERVATION	179
PROCEDURE	651
SYMPTOM	222

Table 3: Number of labeled entities per type in the human-labeled documents.

5 Results

5.1 Corpus Evaluation

Table 4 shows the translation evaluation metrics. BERTScore and LaBSe show quite high correlations between the source and the hypothesis: 0.90 and 0.89, respectively. The KG-UMLS F1 is 0.60, which relies on correct entity linking to UMLS. However, entity translation and linking may also introduce errors, so we calculated an adjusted score, KG-Semantic, that uses a semantic alignment strategy. The KG-Semantic F1 score is 0.87, indicating

[15] https://arunmozhi.in/ner-annotator/

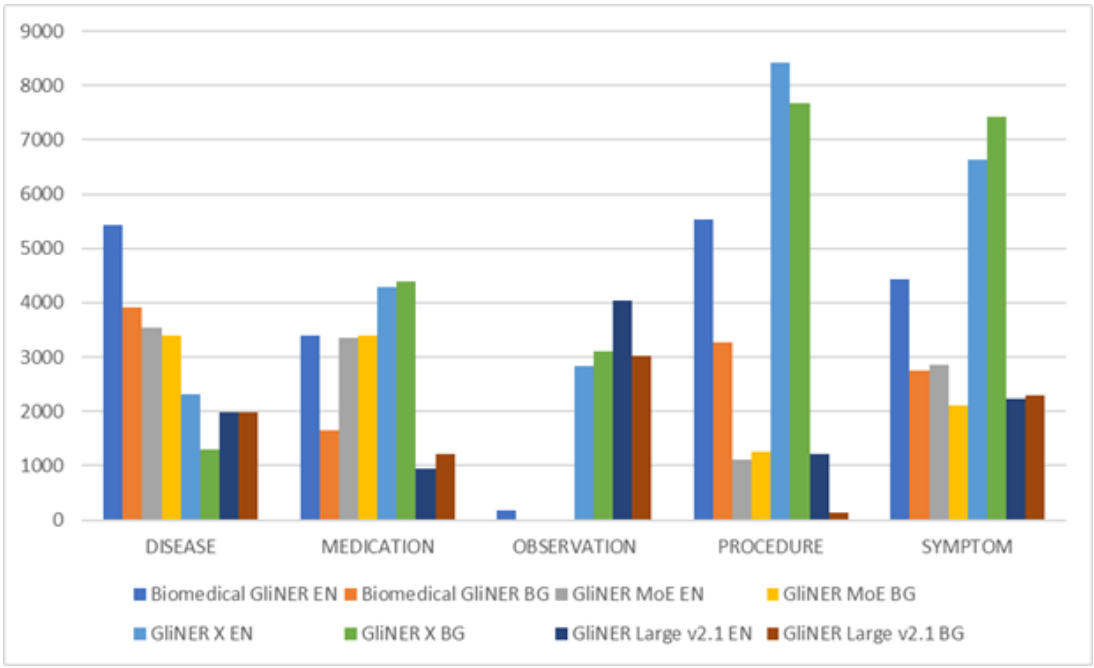

Figure 2: Statistics by type of the number of extracted entities across the original and translated datasets using different models.

good entity-level alignment between the source and translated texts. The combined KG-BERTScore-UMLS for the translation was 0.75, while the KG-BERTScore-Semantic was 0.88.

Metric	Value
BERTScore Precision	0.90
BERTScore Recall	0.90
BERTScore F1	0.90
LaBSe	0.89
KG-UMLS Precision	0.64
KG-UMLS Recall	0.58
KG-UMLS F1	0.60
KG-Semantic Precision	0.89
KG-Semantic Recall	0.87
KG-Semantic F1	0.87
KG-BERTScore-UMLS F1	0.75
KG-BERTScore-Semantic F1	0.88

Table 4: Evaluation metrics for the automatic machine translation.

5.2 Annotation Evaluation

The NER evaluation results are presented in Table 5. The Biomedical GLiNER model achieves the highest F1 score. Both the biomedical and MoE models show higher precision, about 0.73-0.75; however, they struggle with recall, which is especially low for the MoE model at 0.23.

The number of extracted entities per language and entity type is shown in Figure 2. The number of entities extracted in the translated text is significantly lower; about 61% of the entities from the English text are identified in the translation. This result is expected, given that this was a zero-shot approach working in a lower-resource language.

Model	Precision	Recall	F1
Biomedical GLiNER	**0.75**	0.49	**0.59**
GLiNER-X-Large	0.60	**0.51**	0.55
GLiNER Large v2.1	0.64	0.25	0.36
GLiNER Multi-MoE	0.74	0.24	0.36

Table 5: NER model evaluation results on a small subset of manually labeled English documents.

Table 6 shows the number and percentage of linked entities in the UMLS when using the UMLS API for Biomedical GLiNER entities.

Entity Type	UMLS Linked	Total	Percent
Disease	4,941	5,426	91.1%
Medication	3,243	3,396	95.5%
Observation	142	185	76.7%
Procedure	4,973	5,531	89.9%
Symptom	4,202	4,443	94.6%
Grand Total	17,501	18,981	92.2%

Table 6: Entity linking statistics using Biomedical GLiNER and UMLS API.

6 Discussion

LLMs have presented remarkable progress in text generation in the last few years. However, the current models still struggle to generate domain-specific content in low-resource languages. The main reason behind this is data scarcity, and the complexity of domain-specific terminology further complicates the issue. For example, an LLM generating medical, legal, or scientific Bulgarian must simultaneously navigate an underrepresented language in its training corpus, specialized terminology, and a scientific style that demands precise, consistent vocabulary. While high-resource languages like English benefit from vast corpora of technical literature, patents, and academic articles that anchor the terminology of a given field in context, low-resource languages offer the model informal corpora such as web text, social media posts, and general-purpose translations, which barely touch the surface of professional usage and terminology.

This huge gap in training resources available in different languages results in the fabrication of a vocabulary from language- and domain-specific terms. LLMs create a fabricated vocabulary that might be called lexo-phantoms: words that are phonologically plausible, grammatically slottable, but terminologically illegitimate. There are three main types of errors: Transliteration drift, Morphological grafting, and Phantom calques. Transliteration drift is when the model defaults to using phonetic renderings of terms from English or other more widely spoken languages, ignoring established equivalent terms in the low-resource language. Phantom calques are separate word-for-word translations of complex terms. Morphological grafting/blending occurs when morphologically rich languages build technical neologisms through precise systems of suffixes/prefixes/stems. Without enough data to learn these rules, the LLMs graft derivational patterns from dominant languages onto local bases, creating forms that native speakers immediately flag as distorted.

For example, in our experiments when attempting to translate "Patient Discharge Summary", some LLMs invented non-existent words and terms in Bulgarian, like *"изхвърлящо становище"* (eject statement), *"дипяжно съобщение за изписване"* ("diplomatic message for discharge"), *"Дипляно съобщение за изписване на пациент"* ("Folded notice for discharge of a patient"), *"Оригинално ескризе описание на пациента"* ("Original escripture description of the patient"), *"Клиничен изпис на пациента"* ("Clinical discharge of the patient"). The English translations indicated in parentheses are examples, taking the closest-sounding word to the Bulgarian equivalent of the fabricated vocabulary. The lexo-phantoms in the examples above are *"дипяжно"* (diplyazhno), *"Дипляно"* (diplyano), *"ескризе"* (eskrize), *"изпис"* (izpis), which are not loanwords in any linguistically legitimate sense. Another example of literal translation of medical terms is "Encouraged regular professional podiatric care due to increased risk of sensory loss or circulatory impairment" translated as *"Насърчен е редовен професионален педикюр поради повишен риск от сензорна загуба или нарушено кръвообращение"* (regular professional pedicures are encouraged due to increased risk of sensory loss or impaired circulation), which completely changes the meaning of the sentence.

7 Conclusion

The paper presented BGSynthMedic - a public parallel corpus of 450 synthetic discharge summaries in English and Bulgarian, which, to our knowledge, is the first such resource for Bulgarian. We provided the generated corpus in Bulgarian, along with automatic annotations of diseases, medications, symptoms, procedures, and observations for the English and Bulgarian summaries. We described the methods for generating translations and silver-standard annotations and evaluated the quality of the resulting corpus. The corpus is shared with the community for future work in MT, NER, or entity linking in the clinical domain in Bulgarian. The pipeline combines LLM-based machine translation (MT) with reference-free evaluation metrics like BERTScore, LaBSE, and KG-BERTScore. The quality of the translation is evaluated to be relatively high, based on a BERTScore of 0.90 and a LaBSE of 0.89. The proposed method helps address the gap between high- and low-resource languages in specialized domains such as clinical settings. In the future, the approach can be applied to other low-resource languages or specialized domains with complex terminology. Also, extending the manual annotations and benchmarking models on the silver labels could help improve the corpus' utility.

Limitations

The translation method applies to languages supported by general-domain LLMs or to those for

which a language-specific LLM exists, such as BgGPT for Bulgarian. The automatic corpus evaluation relies on zero-shot NER and linking methods; thus, lower scores may stem from NER/linking rather than the translation itself. Also, the evaluation of NER and entity linking was performed on a small subset of the dataset and can be improved to increase the evaluation coverage.

Data Availability

The BGSynthMedic parallel corpus, including synthetic discharge summaries, automatic NER annotations, UMLS entity links, and parallel term and abbreviation dictionaries, is publicly available on GitHub[16], HuggingFace,[17] and Zenodo.[18] The code for generating the parallel corpus is also available on GitHub.[19]

Acknowledgments

This work was partially supported by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria [Grant Project No. BG-RRP-2.004-0008]. The work is also partially supported by the project UNITe BG16RFPR002-1.014-0004 funded by PRIDST. Views and opinions expressed in this paper are of the author only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly to check grammar and spelling. After using these tools/ services, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

Chukwuebuka Anyaegbuna, Eduardo Juan Perez Guerrero, Jerry Liu, Timothy Keyes, April Liang, Natasha Steele, Stephen Ma, Jonathan Chen, and Kevin Schulman. 2026. Multi-method validation of large language model medical translation across high- and low-resource languages.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. MedMT5: An open-source multilingual text-to-text LLM for the medical domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177, Torino, Italia. ELRA and ICCL.

Georgi Grazhdanski, Vasil Vasilev, Sylvia Vassileva, Dimitar Taskov, Izabel Antova, Ivan Koychev, and Svetla Boytcheva. 2025. Synthmedic: Utilizing large language models for synthetic discharge summary generation, correction and validation. *Journal of Biomedical Informatics*, 170:104906.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco-Sánchez, Jan Rodríguez-Miret, and Martin Krallinger. 2023. Overview of symptemist at biocreative viii: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*, page 11.

Salvador Lima-López, Eulàlia Farré-Maduell, Jan Rodríguez-Miret, Miguel Rodríguez-Ortega, Livia Lilli, Jacopo Lenkowicz, Giovanna Ceroni, Jonathan Kossof, Anoop Shah, Anastasios Nentidis, et al. 2024. Overview of multicardioner task at bioasq 2024 on medical specialty and language adaptation of clinical ner systems for spanish, english and italian. In *CLEF (Working Notes)*, pages 8–27.

[16] https://github.com/svassileva/bgsynthmedic

[17] https://huggingface.co/datasets/SU-FMI-AI/BGSynthMedic

[18] https://doi.org/10.5281/zenodo.21333705

[19] https://github.com/svassileva/bg_discharge_summaries

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of ACL-IJCNLP 2021*.

Qiuhao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. 2025. Large language models struggle in token-level clinical named entity recognition. In *AMIA Annual Symposium Proceedings*, volume 2024, page 748.

Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. *CLEF (Working Notes)*, 3180:179–203.

Masoud Monajatipoor, Jiaxin Yang, Joel Stremmel, Melika Emami, Fazlolah Mohaghegh, Mozhdeh Rouhsedaghat, and Kai-Wei Chang. 2024. Llms in biomedicine: A study on clinical named entity recognition. *arXiv preprint arXiv:2404.07376*.

Vandan Mujadia, Pruthwik Mishra, Arafat Ahsan, and Dipti M Sharma. 2023. Towards large language model driven reference-less translation evaluation for english and indian language. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 357–369.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of biomedical semantics*, 9(1):12.

Diana Sousa, Andre Lamurias, and Francisco M. Couto. 2019. A silver standard corpus of human phenotype-gene relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1487–1492, Minneapolis, Minnesota. Association for Computational Linguistics.

Ihor Stepanov and Mykhailo Shtopko. 2024. Gliner multi-task: Generalist lightweight model for various information extraction tasks.

Zhanglin Wu, Min Zhang, Ming Zhu, Yinglu Li, Ting Zhu, Hao Yang, Song Peng, and Ying Qin. 2022. Kg-bertscore: Incorporating knowledge graph into bertscore for reference-free machine translation evaluation. In *Proceedings of the 11th International Joint Conference on Knowledge Graphs*, pages 121–125.

Anthony Yazdani, Ihor Stepanov, and Douglas Teodoro. 2025. Gliner-biomed: A suite of efficient models for open biomedical named entity recognition. *arXiv preprint arXiv:2504.00676*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. Gliner: Generalist model for named entity recognition using bidirectional transformer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix A Category Mapping to UMLS

The mapping of target categories to UMLS semantic types is shown in Table 7.

Entity label	TUI	Semantic type name
disease	T020	Acquired Abnormality
disease	T047	Disease or Syndrome
disease	T048	Mental or Behavioral Dysfunction
disease	T191	Neoplastic Process
medication	T109	Organic Chemical
medication	T121	Pharmacologic Substance
medication	T195	Antibiotic
observation	T033	Finding
observation	T034	Laboratory or Test Result
observation	T184	Sign or Symptom
procedure	T058	Health Care Activity
procedure	T059	Laboratory Procedure
procedure	T060	Diagnostic Procedure
procedure	T061	Therapeutic or Preventive Procedure
symptom	T033	Finding
symptom	T184	Sign or Symptom

Table 7: UMLS semantic types associated with each entity label.

Appendix B Entity Translation Prompt

Prompt BgGPT to translate entities into Bulgarian for entity linking.

Prompt for translation in Bulgarian for clinical entities

You are a medical translator. You will receive a JSON object with a key entities containing a list of Bulgarian medical entity strings. Translate each string to English. Return a JSON object with a single key, translations, whose value is a list of English translations in the same order and with the same count as the input list. Output only valid JSON, without any additional commentary.

Appendix C UMLS Semantic Types mapped to entity labels

Figure 3 summarizes the UMLS semantic type groups and their entity-mapping frequencies. A full list is available in the GitHub repository.

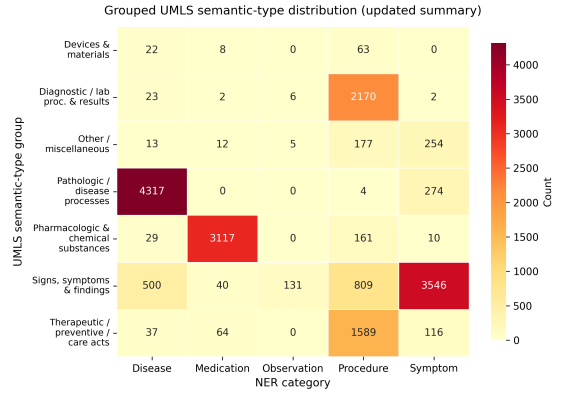

Figure 3: Heatmap of UMLS semantic-type groups across NER categories. Darker cells indicate higher counts.

Appendix D Translation Prompt

Prompt BgGPT to translate a Synthetic Discharge Summary into Bulgarian.

Prompt for translation in Bulgarian of clinical text

You are a highly skilled Natural Language Processing (NLP) expert specializing in medical translation. Your task is to accurately translate English discharge summaries into Bulgarian, ensuring the translated document maintains the original's factual accuracy, uses appropriate medical terminology, correctly interprets and translates abbreviations, and adopts an impersonal style suitable for communication between a hospital and a general practitioner (not directly to the patient). In the "Patient instructions" section, follow an indirect, non-personal style. For example, instead of "Observe", "Apply", "Continue", use "To be observed", "To be applied", "To be continued".
To ensure the highest quality translation, follow this structured approach:
— 1. Terminology Review:
* Identify key medical terms in the English summary.
* Provide the corresponding Bulgarian medical term for each.
* Justify your choice of Bulgarian term, considering context and common usage in Bulgarian medical practice.
Example:
English Term	Bulgarian Term	Justification
—	—	—
Myocardial Infarction	Миокарден инфаркт	Standard Bulgarian medical term for heart attack.
Hypertension	Хипертония	Commonly used and understood term for high blood pressure.
Patient Discharge Summary	Епикриза	Standard Bulgarian medical term for clinical document.
Patient Discharge Summary	Епикриза	Standard Bulgarian medical term for clinical document.
Physical Examination	Обективно състояние / статус	Standard Bulgarian medical term for patient status
Follow-Up and Recommendations	Препоръки и заключение	Standard Bulgarian medical term for patient instructions and recommendations
Treatment and Management	Лечебно-диагностичен план и ход на лечението	Standard Bulgarian medical term for patient treatment during the hospital stay
Hospital Course	Лечебно-диагностичен план и ход на лечението	Standard Bulgarian medical term for patient treatment during the hospital stay
Discharge Instructions	Препоръки и заключение	Standard Bulgarian medical term for patient instructions and recommendations
Chief Complaint	Основни оплаквания	Standard Bulgarian medical term for patient complaint
2. Abbreviation Check:
* List all abbreviations used in the English discharge summary.
* Provide the full English expansion of each abbreviation.
* Provide the corresponding Bulgarian abbreviation (if one exists) and its full Bulgarian expansion. If a direct abbreviation doesn't exist, provide a clear and concise Bulgarian translation of the full term.
Example:
English Abbreviation	English Expansion	Bulgarian Abbreviation	Bulgarian Expansion
—	—	—	—
MI	Myocardial Infarction	МИ	Миокарден инфаркт
BP	Blood Pressure	Артериално кръвно налягане	Кръвно налягане (standard abbreviation АН)
3. Bulgarian Translation:
[Your final, revised Bulgarian translation, incorporating the terminology and abbreviation reviews. Ensure the translation is grammatically correct, fluent, and maintains the original meaning.]
4. Quality Assurance Checklist:
* [ ] Accuracy: Does the Bulgarian translation accurately reflect all information in the English summary?
* [ ] Terminology: Is the medical terminology accurate and appropriate for a Bulgarian medical professional?
* [ ] Abbreviations: Are all abbreviations correctly translated or explained?
* [ ] Grammar: Is the Bulgarian translation grammatically correct and fluent?
* [ ] Clarity: Is the translation clear and easy to understand for a Bulgarian-speaking healthcare professional?
— Now, translate the following discharge summary from English to Bulgarian, adhering to the structure and guidelines outlined above: [Paste the English discharge summary here]."