UNIVERSITY OF CAMBRIDGE

Biomedical Named Entity Recognition via Dictionary-based Synonym Generalization

Zihao Fu[1], Yixuan Su[2], Zaiqiao Meng[3], Nigel Collier[1]

[1]University of Cambridge [2]Cohere
[3]University of Glasgow

EMNLP 2023 https://github.com/fuzihaofzh/BioNER-SynGen

Table of Contents

Introduction

Synonym Generalization Framework

Theoretical Analysis

Experiments

Introduction

- Biomedical Named Entity Recognition (BioNER) recognizes named entities from given text.
- **Example:** "In a study on Homo sapiens, the effect of aspirin on heart disease was investigated."
  - Homo sapiens, aspirin, and heart disease are biomedical entities.

Introduction

- Existing main approaches:
  1. **Supervised methods:** Require large-scale human-annotated data for training. (Wang et al. (2019b); Lee et al. (2020); Weber et al. (2021))
  2. **Distantly supervised methods:** Use weakly annotated data based on an in-domain training corpus. (Fries et al. (2017); Zhang et al. (2021); Zhou et al. (2022))
  3. **Dictionary-based methods:** Trained with predefined dictionaries. (Aronson (2001); Song et al. (2015); Soldaini and Goharian (2016); Nayel et al. (2019); Basaldella et al. (2020))
- Focus: Dictionary-based method for BioNER.

Dictionary-based Approaches

Advantages:

- No need for human-annotated data or in-domain corpus.
- Popular choice due to low human effort and expert involvement.

Challenges:

Suffers from the *synonym generalization problem*:

- Limited to entities present in the dictionary.
- Cannot recognize synonyms outside the dictionary.
- Existing techniques like string similarity are not enough.

Contributions

- **SynGen Framework:** A novel dictionary-based method for BioNER, addressing the synonym generalization problem.
- **Theoretical Analysis:** Demonstrated that SynGen optimization is equivalent to minimizing the synonym generalization error.
- **Extensive Experiments:** Validated SynGen's effectiveness over various benchmarks, surpassing prior dictionary-based models.

Table of Contents

Introduction

Synonym Generalization Framework

Theoretical Analysis

Experiments

Synonym Generalization Framework

- Training stage: SynGen samples synonyms and learns classification.
- Inference stage: Split input text and score spans.

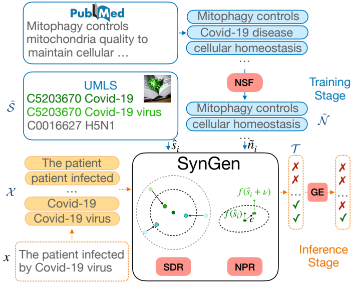


Figure: SynGen framework. → represents the training steps while - - → represents the inference steps.

Task Definition

- Biomedical domain denoted as $\mathcal{D}$.
- All possible biomedical entities in $\mathcal{D}$: $\mathcal{S} = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_{|\mathcal{S}|}\}$.
- Task: Identify sub-spans in text $\boldsymbol{x}$ that belong to $\mathcal{S}$. (Find $\{\boldsymbol{x}_{[b_1:e_1]}, \cdots, \boldsymbol{x}_{[b_k:e_k]} | \forall i \in [1, k], \boldsymbol{x}_{[b_i:e_i]} \in \mathcal{S}\}$)
- Real-world constraint: Only access to subset $\hat{\mathcal{S}} \subset \mathcal{S}$ from a dictionary.

Training Stage

- Sample synonyms as positive samples and spans from biomedical corpus (e.g. PubMed) as negative samples.
- Objective: Classify samples using cross-entropy.
- Regularizations introduced: Synonym distance regularization and noise perturbation regularization.

Dictionary-Based Loss

- Sample positive sample $\hat{\boldsymbol{s}}_i$ from dictionary $\hat{\mathcal{S}}$ is encoded.
- Encoding of the sample: $\hat{\boldsymbol{r}}_i = E(\hat{\boldsymbol{s}}_i)$.
- Probability of $\hat{\boldsymbol{s}}_i$ being a span of entity: $p(\hat{\boldsymbol{s}}_i) = \sigma(\mathrm{MLP}(\hat{\boldsymbol{r}}_i))$.
- Negative sample $\tilde{\boldsymbol{n}}_i$ from PubMed with similar computation.
- Dictionary-based loss: $\mathcal{L}_c = -\frac{1}{2|\hat{\mathcal{S}}|}\sum_{i=0}^{|\hat{\mathcal{S}}|} \ln p(\hat{\boldsymbol{s}}_i) + \ln[1 - p(\tilde{\boldsymbol{n}}_i)]$.

Negative Sampling Filtering (NSF)

To obtain negative sample $\tilde{\boldsymbol{n}}_i$:

- Sample random-length spans from the PubMed corpus.
- Remove samples close to dictionary entities.
- Ensures:

$$\min_{\forall \hat{\boldsymbol{s}}_i \in \hat{\mathcal{S}}, \forall \tilde{\boldsymbol{n}}_j \in \tilde{\mathcal{N}}} \| F(\hat{\boldsymbol{s}}_i) - F(\tilde{\boldsymbol{n}}_j) \| > t_d$$

where t_d: Threshold of minimal distance.

Synonym Distance Regularizer (SDR)

- Goal: Equip the model to identify more synonyms of the same biomedical concept. If the embeddings of these synonyms are close to each other, it helps in correct identification.
- Procedure:
  - Sample anchor entity $\hat{\boldsymbol{s}}_a$ and its synonym $\hat{\boldsymbol{s}}_p$ from the dictionary $\hat{\mathcal{S}}$. They share the same concept ID.
  - Sample a random negative $\tilde{\boldsymbol{n}}_n \in \tilde{\mathcal{N}}$.
  - Impose a triplet margin loss:

$$\mathcal{R}_s = \max\{\|\hat{\boldsymbol{r}}_a - \hat{\boldsymbol{r}}_p\| - \|\hat{\boldsymbol{r}}_a - \tilde{\boldsymbol{r}}_n\| + \gamma_s, 0\}$$

where:

  - γ_s: Pre-defined margin.
  - $\hat{\boldsymbol{r}}_a = E(\hat{\boldsymbol{s}}_a)$, $\hat{\boldsymbol{r}}_p = E(\hat{\boldsymbol{s}}_p)$, and $\tilde{\boldsymbol{r}}_n = E(\tilde{\boldsymbol{n}}_n)$.

Noise Perturbation Regularizer (NPR)

- Goal: Reduce the sharpness of the scoring function's landscape to give close entities similar scores. Synonyms of a biomedical entity are expected to be close.
- NPR Definition:

$$\mathcal{R}_n = \|p(\hat{\boldsymbol{r}}_i + \boldsymbol{v}) - p(\hat{\boldsymbol{r}}_i)\|$$

  where:

  - $\hat{\boldsymbol{r}}_i$: Embedding of biomedical entity sampled from $\hat{\mathcal{S}}$.
  - $\boldsymbol{v}$: Gaussian noise vector.
- NPR aims to flatten the landscape of the loss function, minimizing the loss difference for vectors within close regions.
- Further discussion on function flatness: Foret et al. (2020); Bahri et al. (2022).

Overall Loss

$$\mathcal{L} = \mathcal{L}_c + \alpha\mathcal{R}_s + \beta\mathcal{R}_n$$

where:

- α, β: Tunable hyperparameters for the regularizers.
- $\mathcal{L}_c$: Cross-entropy objective.
- $\mathcal{R}_s$: Synonym Distance Regularizer.
- $\mathcal{R}_n$: Noise Perturbation Regularizer.

Inference

- Split input text $\boldsymbol{x}$ into spans:

$$\mathcal{X} = \{\boldsymbol{x}_{[i:j]} | 0 \leq i \leq j \leq |\boldsymbol{x}|, j - i \leq m_s\}$$

- Score each span:

$$\text{score}(\boldsymbol{x}_{[i:j]}) = \sigma(\text{MLP}(E(\boldsymbol{x}_{[i:j]})))$$

  Only retain spans where $\text{score}(\boldsymbol{x}_{[i:j]}) > t$.
- Apply Greedy Extraction (GE) to extract biomedical terms.

Greedy Extraction

- It's observed that biomedical terms, such as *T-cell prolymphocytic leukemia*, can be nested, containing sub-entities like *T-cell* and *leukemia* Finkel and Manning (2009); Marinho et al. (2019).
- SynGen's GE ranks recognized terms by length in descending order:

  $$\mathcal{T} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n | \forall i < j, |\boldsymbol{t}_i| > |\boldsymbol{t}_j|\}$$

  Start with the initial validation sequence $\boldsymbol{x}^{(1)} = \boldsymbol{x}$.
- For each term $\boldsymbol{t}_i$ in $\mathcal{T}$:
  - If $\boldsymbol{t}_i$ is a sub-sequence of the current validation sequence $\boldsymbol{x}^{(i)}$, i.e., $\exists p, q < |\boldsymbol{x}^{(i)}|$ such that $\boldsymbol{t}_i = \boldsymbol{x}^{(i)}_{[p:q]}$, recognize $\boldsymbol{t}_i$ as a biomedical entity.
  - Update the validation sequence: $\boldsymbol{x}^{(i+1)}$ removes all occurrences of $\boldsymbol{t}_i$ in $\boldsymbol{x}^{(i)}$.

Table of Contents

Introduction

Synonym Generalization Framework

Theoretical Analysis

Experiments

Theoretical Analysis

Most existing dictionary-based frameworks struggle with the *synonyms generalization problem*, where terms outside the dictionary aren't easily recognized.

Objective:

- Address the *synonyms generalization problem* with the SynGen framework.

Analysis Methodology:

- Measure the average empirical error for entities in the dictionary $\hat{\mathcal{S}}$.
- Define the synonym generalization error using the most pessimistic error gap.

Loss Function and Average Empirical Error

- **Loss Function:** The bounded negative log-likelihood measures the correctness of classification. A value of 0 indicates a correct classification while a value of b indicates a misclassification.

$$f(\boldsymbol{r}) = -\ln \sigma(\mathrm{MLP}(\boldsymbol{r})) \in [0, b], \quad \boldsymbol{r} = E(\boldsymbol{s})$$

- **Average Empirical Error:** It represents the average of the loss function values across all entities in the dictionary $\hat{\mathcal{S}}$. A lower value indicates a better generalization performance.

$$\hat{R} = \frac{1}{|\hat{\mathcal{S}}|} \sum_{i=1}^{|\hat{\mathcal{S}}|} f(\hat{\boldsymbol{r}}_i)$$

- **Implication:** If $\hat{R}$ is low, the model performs well on dictionary terms. How about the performance of other biomedical entities outside the dictionary?

Synonym Generalization Error

- In dictionary-based frameworks, how well does the model generalize to synonyms not present in the dictionary? The “Synonym Generalization Error” quantifies this.

Definition (synonym generalization error)

Given a loss function $f(\boldsymbol{r}) \in [0, b]$:

$$E_s = \sup_{\boldsymbol{s} \in \mathcal{S}} (f(E(\boldsymbol{s})) - \hat{R})$$

- **Interpretation:**
  - Small E_s implies the error for any $\boldsymbol{s}$ will be close to $\hat{R}$.
  - High E_s indicates the model may struggle with unseen synonyms.
- **Implication:** A low E_s suggests that training with the dictionary terms $\hat{\mathcal{S}}$ will generalize well to other biomedical entities in the domain $\mathcal{S}$.

Assumptions for Analysis

- **Entity Cover**:
  - $\hat{\mathcal{S}}$ is an $\epsilon-$net of $\mathcal{S}$.
  - Implication: For any entity in $\mathcal{S}$, there's a close representative in the sampled dictionary $\hat{\mathcal{S}}$.

$$\forall \boldsymbol{s} \in \mathcal{S}, \exists \hat{\boldsymbol{s}} \in \hat{\mathcal{S}}, \|\hat{\boldsymbol{s}} - \boldsymbol{s}\| \leq \epsilon$$

- **Flatness of Loss Function**:
  - f is κ-Lipschitz.
  - Implication: The loss function doesn't change too rapidly between close points.

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \leq \kappa \|\boldsymbol{x} - \boldsymbol{y}\|$$

Synonym Generalization Error Bound

Theorem (Synonym Generalization Error Bound)

Given the assumptions, with probability at least $1 - \delta$*:*

$$E_s < (\kappa\epsilon + b)\sqrt{\frac{\ln|\mathcal{S}| + \ln\frac{2}{\delta}}{2}} + b\sqrt{\frac{\ln\frac{2}{\delta}}{2|\hat{\mathcal{S}}|}}$$

- **Dictionary Density (ϵ):** Refers to how tightly entities in $\mathcal{S}$ are clustered around those in $\hat{\mathcal{S}}$. Smaller ϵ ensures better generalization.
- **Lipschitz Constant (κ):** Represents stability of the loss function. Smaller κ ensures the function doesn't have abrupt changes, leading to model robustness.
- **Dictionary Size ($|\hat{\mathcal{S}}|$):** Larger $|\hat{\mathcal{S}}|$ helps to reduce the error bound leading to better generalization. It also explains why it supports few-shot training.
- **All Terms Count ($|\mathcal{S}|$):** Larger $|\mathcal{S}|$ leads to bad generalization bound. Luckily, it get worse at the rate of $\mathcal{O}(\sqrt{\ln|\mathcal{S}|})$.

Implications of The Bound - SDR

Theorem (Synonym Generalization Error Bound)

With certain assumptions, with probability at least $1-\delta$*:*

$$E_s < (\kappa\epsilon + b)\sqrt{\frac{\ln|\mathcal{S}| + \ln\frac{2}{\delta}}{2}} + b\sqrt{\frac{\ln\frac{2}{\delta}}{2|\hat{\mathcal{S}}|}}$$

SDR Reduces Synonym Distance:

- SDR aims to decrease the distance between synonyms.
- This reduction is analogous to minimizing ϵ.
- Consequently, it diminishes the synonym generalization error upper bound.

Implications of The Bound - NPR

Theorem (Synonym Generalization Error Bound)

Given the assumptions, with probability at least $1 - \delta$*:*

$$E_s < (\kappa\epsilon + b)\sqrt{\frac{\ln|\mathcal{S}| + \ln\frac{2}{\delta}}{2}} + b\sqrt{\frac{\ln\frac{2}{\delta}}{2|\hat{\mathcal{S}}|}}$$

NPR Reduces Lipschitz Constant:

- Lipschitz constant κ is calculated as:

$$\frac{\|f(\hat{\boldsymbol{x}}_i + \boldsymbol{v}) - f(\hat{\boldsymbol{x}}_i)\|}{\|(\hat{\boldsymbol{x}}_i + \boldsymbol{v}) - \hat{\boldsymbol{x}}_i\|}$$

- Minimizing $\mathcal{R}_n$ is equivalent to minimizing κ.

Table of Contents

Introduction

Synonym Generalization Framework

Theoretical Analysis

Experiments

Experimental Setup

- Evaluated on 6 popular BioNER datasets: BC2GM, BC4CHEMD, BC5CDR, JNLPBA, NCBI-Disease, S800.
- Performance metrics: Precision (P), Recall (R), and F_1 scores.
- Entity name dictionary: concepts' synonyms from UMLS.
- Hyper-parameters tuned using grid search.
- Negative spans from PubMed corpus.
- Backbone model: PubMedBert.
- Tests run on NVIDIA GeForce RTX 3090 GPUs.

Baseline Models

(Distantly) Supervised:

- **BioBert** Lee et al. (2020) first pre-trains an encoder with biomedical corpus and then fine-tunes the model on annotated NER datasets.
- **SBM** is a standard Span-Based Model Lee et al. (2017); Luan et al. (2018, 2019); Zhong and Chen (2021) for NER task.
- **SBMCross** utilizes the same model as SBM. Train and test on different dataset in the same domain Langnickel and Fluck (2021).
- **SWELLSHARK** Fries et al. (2017) uses weak supervision.
- **AutoNER** Wang et al. (2019a); Shang et al. (2020) trains AutoPhase Shang et al. (2018) and tailors dictionary.

Baseline Models

Dictionary-Based:

- **EmbSim** uses pre-trained model for encoding.
- **MetaMap** Aronson (2001); Divita et al. (2014); Soldaini and Goharian (2016) does exact concept mapping.
- **SPED** Rudniy et al. (2012); Song et al. (2015) calculates Shortest Path Edit Distances.
- **TF-IDF** follows the model of Ujiie et al. (2021) as the similarity score.
- **QuickUMLS** Soldaini and Goharian (2016) uses Simstring Okazaki and Tsujii (2010) to extract entities.

Main Results

	Model	NCBI			BC5CDR-D			BC5CDR-C			BC4CHEMD			Species-800			LINNAEUS			AVG		
		P	R	F$_1$	P	R	F$_1$	P	R	F$_1$	P	R	F$_1$	P	R	F$_1$	P	R	F$_1$	P	R	F$_1$
(Distantly) Supervised	BioBert$^\natural$	88.2	91.2	89.7	86.5	87.8	87.2	93.7	93.3	93.5	92.8	91.9	92.4	72.8	75.4	74.1	90.8	85.8	88.2	87.5	87.6	87.5
	SBM$^\natural$	88.4	88.9	88.6	83.4	86.4	84.9	93.2	93.6	93.4	92.0	86.6	89.2	99.5	91.6	95.4	99.8	80.1	88.9	92.8	87.9	90.1
	SBMCross$^\diamondsuit$	75.9	58.3	66.0	70.1	61.3	65.4	94.1	86.4	90.1	72.2	63.2	67.4	64.2	64.5	64.3	78.8	45.8	57.9	75.9	63.2	68.5
	SWELLSHARK$^{\flat\triangle}$	64.7	69.7	67.1	80.7	77.6	79.1	88.3	88.3	88.3	-	-	-	-	-	-	-	-	-	77.9	78.5	78.2
	AutoNER$^{\heartsuit\sharp}$	79.4	72.0	75.5	86.2	67.9	76.0	85.2	84.2	84.7	91.1	18.9	31.3	86.6	90.9	88.7	92.1	95.6	93.8	86.8	71.6	75.0
	AutoNER w/o DT$^\heartsuit$	66.8	32.4	43.6	72.0	17.3	27.9	89.7	67.3	76.9	90.7	19.7	32.4	57.6	50.7	53.9	88.4	39.0	54.1	77.5	37.7	48.1
	AutoNER w/o IDC$^\sharp$	85.1	19.1	31.2	87.1	40.4	55.2	94.2	37.3	53.4	91.2	18.8	31.2	83.6	18.5	30.3	90.4	62.8	74.1	88.6	32.8	45.9
Dictionary-Based	AutoNER w/o DT+IDC	57.9	9.7	16.6	63.0	13.9	22.8	92.8	39.3	55.2	60.9	24.6	35.1	59.8	25.0	35.3	80.1	33.0	46.8	69.1	24.2	35.3
	EmbSim	56.7	24.9	34.6	61.8	14.3	23.2	71.7	61.2	66.0	47.4	24.7	32.4	49.0	34.2	40.3	80.4	42.9	55.9	61.2	33.7	42.1
	MetaMap	61.8	27.8	38.4	69.3	13.3	22.3	65.9	63.5	64.7	33.1	25.2	28.6	56.9	48.7	52.5	85.5	44.3	58.3	62.1	37.1	44.1
	MetaMap (Uncased)	58.4	27.5	37.4	63.5	18.4	28.6	**94.8**	64.1	76.5	**86.2**	24.0	37.5	49.1	52.3	50.6	79.1	49.6	61.0	71.9	39.3	48.6
	SPED	59.3	30.1	39.9	68.2	14.3	23.7	65.6	63.9	64.8	33.0	25.4	28.7	56.0	49.4	52.5	85.3	44.7	58.7	61.2	38.0	44.7
	TF-IDF	26.1	29.7	27.7	32.0	22.6	26.4	74.1	65.4	69.5	19.1	39.3	25.7	42.5	21.4	28.4	77.3	40.5	53.1	45.2	36.5	38.5
	QuickUMLS	**80.4**	17.2	28.4	**93.5**	14.5	25.1	93.2	56.9	70.7	82.7	16.9	28.1	**61.7**	46.7	53.2	**88.2**	44.7	59.3	**83.3**	32.8	44.1
	SynGen	68.8	**64.1**	**66.2**	63.8	**63.4**	**63.5**	85.0	**83.9**	**84.4**	56.4	**51.1**	**53.6**	58.8	**65.7**	**62.0**	84.9	**66.2**	**74.4**	69.6	**65.7**	**67.4**

Main Results:

- (1) SynGen outperforms other models in F_1 score, improving recall by capturing more entities.
- (2) By comparing SBM and SBMCross,performance varies with the choice of in-domain corpus. Wrong choice can decrease performance.
- (3) SynGen's F_1 scores (67.4) are close to SBMCross (68.5), showing its effectiveness.

Main Results

	Model	NCBI			BC5CDR-D			BC5CDR-C			BC4CHEMD			Species-800			LINNAEUS			AVG		
		P	R	F$_1$	P	R	F$_1$	P	R	F$_1$	P	R	F$_1$	P	R	F$_1$	P	R	F$_1$	P	R	F$_1$
(Distantly) Supervised	BioBert$^\natural$	88.2	91.2	89.7	86.5	87.8	87.2	93.7	93.3	93.5	92.8	91.9	92.4	72.8	75.4	74.1	90.8	85.8	88.2	87.5	87.6	87.5
	SBM$^\natural$	88.4	88.9	88.6	83.4	86.4	84.9	93.2	93.6	93.4	92.0	86.6	89.2	99.5	91.6	95.4	99.8	80.1	88.9	92.8	87.9	90.1
	SBMCross$^\diamond$	75.9	58.3	66.0	70.1	61.3	65.4	94.1	86.4	90.1	72.2	63.2	67.4	64.2	64.5	64.3	78.8	45.8	57.9	75.9	63.2	68.5
	SWELLSHARK$^{\flat\triangle}$	64.7	69.7	67.1	80.7	77.6	79.1	88.3	88.3	88.3	-	-	-	-	-	-	-	-	-	77.9	78.5	78.2
	AutoNER$^{\heartsuit\sharp}$	79.4	72.0	75.5	86.2	67.9	76.0	85.2	84.2	84.7	91.1	18.9	31.3	86.6	90.9	88.7	92.1	95.6	93.8	86.8	71.6	75.0
	AutoNER w/o DT$^\heartsuit$	66.8	32.4	43.6	72.0	17.3	27.9	89.7	67.3	76.9	90.7	19.7	32.4	57.6	50.7	53.9	88.4	39.0	54.1	77.5	37.7	48.1
	AutoNER w/o IDC$^\sharp$	85.1	19.1	31.2	87.1	40.4	55.2	94.2	37.3	53.4	91.2	18.8	31.2	83.6	18.5	30.3	90.4	62.8	74.1	88.6	32.8	45.9
Dictionary-Based	AutoNER w/o DT+IDC	57.9	9.7	16.6	63.0	13.9	22.8	92.8	39.3	55.2	60.9	24.6	35.1	59.8	25.0	35.3	80.1	33.0	46.8	69.1	24.2	35.3
	EmbSim	56.7	24.9	34.6	61.8	14.3	23.2	71.7	61.2	66.0	47.4	24.7	32.4	49.0	34.2	40.3	80.4	42.9	55.9	61.2	33.7	42.1
	MetaMap	61.8	27.8	38.4	69.3	13.3	22.3	65.9	63.5	64.7	33.1	25.2	28.6	56.9	48.7	52.5	85.5	44.3	58.3	62.1	37.1	44.1
	MetaMap (Uncased)	58.4	27.5	37.4	63.5	18.4	28.6	**94.8**	64.1	76.5	**86.2**	24.0	37.5	49.1	52.3	50.6	79.1	49.6	61.0	71.9	39.3	48.6
	SPED	59.3	30.1	39.9	68.2	14.3	23.7	65.6	63.9	64.8	33.0	25.4	28.7	56.0	49.4	52.5	85.3	44.7	58.7	61.2	38.0	44.7
	TF-IDF	26.1	29.7	27.7	32.0	22.6	26.4	74.1	65.4	69.5	19.1	39.3	25.7	42.5	21.4	28.4	77.3	40.5	53.1	45.2	36.5	38.5
	QuickUMLS	**80.4**	17.2	28.4	**93.5**	14.5	25.1	93.2	56.9	70.7	82.7	16.9	28.1	**61.7**	46.7	53.2	**88.2**	44.7	59.3	**83.3**	32.8	44.1
	SynGen	68.8	**64.1**	**66.2**	63.8	**63.4**	**63.5**	85.0	**83.9**	**84.4**	56.4	**51.1**	**53.6**	58.8	**65.7**	**62.0**	84.9	**66.2**	**74.4**	69.6	**65.7**	**67.4**

Main Results:

- (4) QuickUMLS has high precision using exact matches but struggles with out-of-dictionary synonyms, affecting recall.
- (5) For AutoNER, custom dictionaries and correct in-domain corpus are crucial. Our model performs better without these specifics.

Ablation Study

	NCBI	BC5CDR-D	BC5CDR-C	BC4C HEMD	Species-800	LINNAEUS	AVG
SynGen	**66.2**	**63.5**	**84.4**	**53.6**	**62.0**	**74.4**	**67.4**
w/o SDR	66.2	63.1	80.8	51.1	60.6	73.0	65.8
w/o NPR	66.2	62.8	78.0	49.7	60.3	73.1	65.0
w/o NPR+SDR	64.7	58.7	76.9	49.0	59.7	72.0	63.5
w/o NSF	60.3	49.5	76.8	49.0	54.3	54.7	57.4
w/o GE	49.8	54.4	69.4	33.4	52.1	71.7	55.1

1. Variants without NPR or SDR drop in performance.
2. Negative sample filtering (NSF) is crucial. Observe SynGen w/o NSF.
3. The Greedy Extraction (GE) model boosts precision. Refer to SynGen w/o GE.

Impact of the SDR Component

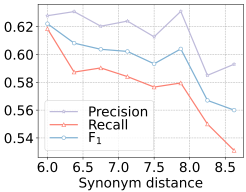


Figure: Influence of synonym distance.

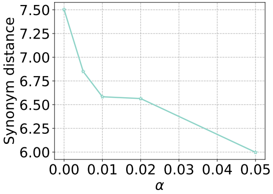


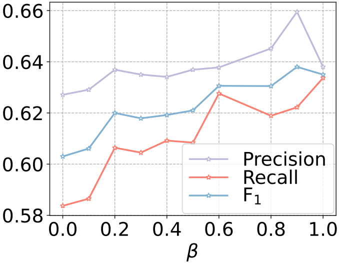
Figure: Influence of synonym distance regularizer's weight α.

Recall loss $\mathcal{L} = \mathcal{L}_c + \alpha\mathcal{R}_s + \beta\mathcal{R}_n$. Observations:

1. Model trained with varied hyper-parameter α.
2. 10,000 synonym pairs sampled from UMLS to measure distance.
3. As α increases, synonym distance decreases. Refer to right figure.
4. Evaluation scores improve as synonym distance is regularized. (left)

Result: SDR component effectively controls synonym distance and boosts performance, supporting the analysis in our theory.

Influence of Noise Perturbation



Recall loss $\mathcal{L} = \mathcal{L}_c + \alpha\mathcal{R}_s + \beta\mathcal{R}_n$. Observations:

1. We investigated score changes with varying NPR weight (β).
2. As β increases, precision, recall, and F_1 scores all rise.
3. The NPR component evidently enhances model performance.
4. This observation supports our theoretical analysis.

Few-Shot Analysis

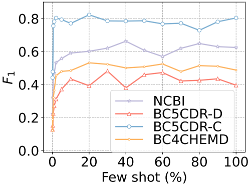


Figure: Few-shot analysis.

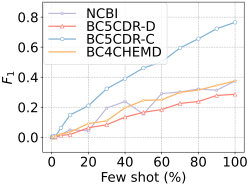


Figure: Few-shot analysis for MetaMap.

1. SynGen performance rises with dictionary size, especially when it's small.
2. Using just 20% of dictionary entries gives results comparable to using the full dictionary. Performance plateaus after a specific dictionary size ratio. (Our theory can explain!)
3. In contrast, MetaMap performance increases linearly with dictionary size.
4. Word match-based models, like MetaMap, aren't suited for few-shot cases.

Standard Deviation Analysis

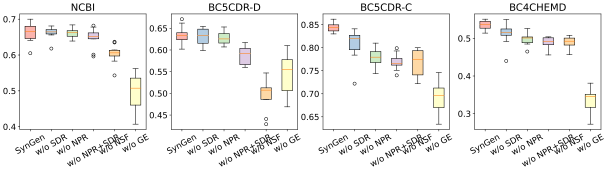


Figure: The box plot of each model's F_1 score over 10 runs.

- Conducted standard deviation analysis using 10 runs with different random seeds.
- SynGen consistently outperforms model variants lacking proposed components.
- Validates both the individual effectiveness and overall consistency of SynGen components.

Conclusion

- Introduced SynGen: a novel synonym generalization framework for the BioNER task with a dictionary.
- Proposed two new regularizers to enhance term generalizability across the domain.
- Performed comprehensive theoretical analysis highlighting the efficacy of the proposed components in dictionary-based biomedical NER tasks.
- Extensive evaluations on multiple benchmarks confirm that SynGen significantly surpasses prior dictionary-based models.

Thank You

Questions?

`https://github.com/fuzihaofzh/BioNER-SynGen`

References I

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Dara Bahri, Hossein Mobahi, and Yi Tay. 2022. Sharpness-aware minimization improves language model generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7360–7371.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. Cometa: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137.

Guy Divita, Qing T Zeng, Adi V Gundlapalli, Scott Duvall, Jonathan Nebeker, and Matthew H Samore. 2014. Sophia: a expedient umls concept extraction annotator. In *AMIA Annual Symposium Proceedings*, volume 2014, page 467. American Medical Informatics Association.

Jenny Rose Finkel and Christopher D Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 141–150.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2020. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*.

Jason Fries, Sen Wu, Alex Ratner, and Christopher Ré. 2017. Swellshark: A generative model for biomedical named entity recognition without labeled data. *arXiv preprint arXiv:1704.06360*.

Lisa Langnickel and Juliane Fluck. 2021. We are not ready yet: limitations of transfer learning for disease named entity recognition. *bioRxiv*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234.

References II

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. 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 3036–3046.

Zita Marinho, Alfonso Mendes, Sebastiao Miranda, and David Nogueira. 2019. Hierarchical nested named entity recognition. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 28–34.

Hamada A Nayel et al. 2019. Integrating dictionary feature into a deep learning model for disease named entity recognition. *arXiv preprint arXiv:1911.01600*.

Naoaki Okazaki and Jun'ichi Tsujii. 2010. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 851–859.

Alex Rudniy, James Geller, and Min Song. 2012. Histogram difference string distance for enhancing ontology integration in bioinformatics. In *4th International Conference on Bioinformatics and Computational Biology 2012, BICoB 2012*, pages 108–113.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825–1837.

References III

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2020. Learning named entity tagger using domain-specific dictionary. In *2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2054–2064. Association for Computational Linguistics.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Min Song, Hwanjo Yu, and Wook-Shin Han. 2015. Developing a hybrid dictionary-based bio-entity recognition technique. *BMC medical informatics and decision making*, 15(1):1–8.

Shogo Ujiie, Hayate Iso, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2021. End-to-end biomedical entity linking with span-based dictionary matching. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 162–167.

Xuan Wang, Yu Zhang, Qi Li, Xiang Ren, Jingbo Shang, and Jiawei Han. 2019a. Distantly supervised biomedical named entity recognition with dictionary expansion. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 496–503. IEEE.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2019b. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35(10):1745–1752.

Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2021. Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794.

Xinghua Zhang, Bowen Yu, Tingwen Liu, Zhenyu Zhang, Jiawei Sheng, Xue Mengge, and Hongbo Xu. 2021. Improving distantly-supervised named entity recognition with self-collaborative denoising learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10746–10757.

References IV

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61.

Kang Zhou, Yuepei Li, and Qi Li. 2022. Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7198–7211.