Fine-Grained Mention-Level Analysis of Biomedical Entity Linking Models

MIE 2026 — Medical Informatics Europe

Baptiste Pras Nona Naderi

May 2026

Université Paris-Saclay, CNRS, LISN — 91400 Orsay, France

baptiste.pras@universite-paris-saclay.fr

Biomedical Entity Linking (BEL)

The Task

- Map mentions to standardized IDs (UMLS, MeSH).
- Essential for: clinical decision support & data mining.

Examples

"Patient is prescribed **Doliprane**."

→ **Paracetamol** (UMLS:C0086787)

"**Dynactin 4** overexpression was observed."

→ **DCTN4 protein, human** (UMLS:C4308010)

Main Challenges

- **Synonymy:** Tylenol ↔ Paracetamol.
- **Morphology:** *a-glucosidase* vs *alpha glucosidase*.
- **Ambiguity:** *"Cold"* (Infection vs Temp).
- **Abbreviations:** BRCA1, p53, HIV.

The Problem: Blind Spots in Evaluation

Limitations of Global Recall@1
Aggregate scores mask failures on:

- Rare entities
- Complex abbreviations
- Zero-shot cases

A model can score 70% overall while being 0% reliable on rare diseases.

Our Contribution

- A **fine-grained** framework.
- Explains *where* and *why* models fail.
- Uses 6 interpretable mention-level characteristics.

A Framework Based on Six Mention Characteristics

Mention difficulty

- **Length**: short ($\leq$10 tokens) vs. long.
  E.g., "fever" vs. "maturity onset diabetes of the young type 7".
- **Lexical variation**: Levenshtein distance from mention to closest KB synonym: low ($\leq$0.1) vs. high ($>$0.1).
- **Synonymy**: number of distinct names for the gold concept in the KB.
- **Homonymy**: surface form maps to $>$1 concept in the KB.
  E.g., "cold" $\rightarrow$ viral infection *or* temperature condition.

Training coverage

- **Mention frequency**: how often the exact mention string appeared in training. $\rightarrow$ frequent / rare ($\leq$10) / zero-shot.
- **Entity frequency**: how often the gold concept appeared in training. $\rightarrow$ frequent / rare / zero-shot.

All characteristics computed **automatically** from the KB and training set.

Experimental Setup

Evaluation: BELB

The **Biomedical Entity Linking Benchmark (BELB)** is a unified framework for fair, reproducible comparison of BEL models: standardized preprocessing, common KBs, consistent metric.

Datasets (3 corpora, diverse domains)

- **Linnaeus**: species in biomedical literature.
- **S800**: organisms in scientific abstracts.
- **MedMentions**: UMLS-annotated abstracts (diseases, drugs, genes, proteins...).

Metric: **recall@1** (top-1 prediction matches gold entity).

Models

ArboEL BERT dual-encoder; scores mention–entity pairs; global consistency via arborescence.

GenBioEL BART generative model; generates entity names token by token; prefix-constrained decoding over KB.

RBES Rule-based, string-matching baseline.

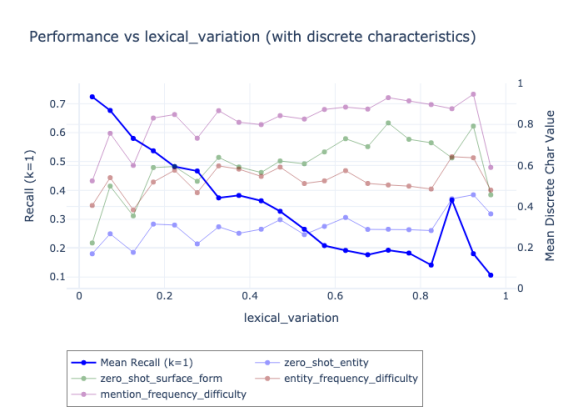
Our framework is applied to the **test mentions** after running all models through BELB.

Results: Systematic Drops Across All Models on All Datasets

Characteristic (difficult case)	GenBioEL	ArboEL	RBES
Zero-shot entity	↓0.13	↓0.27	↓0.05
Zero-shot name	↓0.16	↓0.31	↓0.19
Long mention	↓0.09	↓0.25	↓0.19
Few synonyms	↓0.15	↓0.10	↓0.14
Homonyms exist	↓0.07	↓0.01	↓0.11
High lexical variation	**↓0.39**	**↓0.42**	**↓0.56**
Rare entity	↓0.08	↓0.17	↓0.02
Rare mention	↓0.08	↓0.23	↓0.12

Drops relative to easy case. All confirmed by narrow 95% bootstrap CI.

The Hardest Challenge: Lexical Variation



High lexical variation: the mention surface form differs substantially from any name in the KB.

Non-standard surface forms are difficult for *all* systems.

Especially frequent for
Gene and protein names: abbreviations, species variants, isoforms. E.g., “a-glucosidase” (text) vs. “alpha glucosidase” (KB name).

Right y-axis: proportion of zero-shot/rare cases at each level of lexical variation.

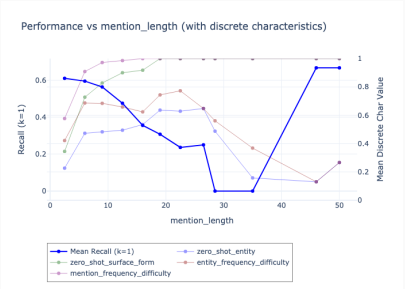
Training Sparsity as the Underlying Cause

Across all characteristics (lexical variation, mention length, synonymy, homonymy):

- As a mention becomes **harder**, the proportion of zero-shot/rare cases **increases**.
- Performance drops consistently reflect limited training coverage.

McNemar test: $p = 3.55 \times 10^{-96}$ (gap is not due to chance).

Key insight
Despite Transformer encoders (BERT/BART), both models rely heavily on **surface form matching** and **entity training frequency**. Contextual generalization to rare or non-standard forms remains limited.



As mention length increases, zero-shot/rare prevalence rises, and performance drops accordingly.

Conclusion

Fine-grained BEL evaluation framework

- 6 interpretable, automatically computed characteristics.
- Applied across 3 datasets, 3 models, via BELB.
- Statistical validation: 95% bootstrap CI.

BELB integration

Our framework extends BELB with interpretable mention-level annotations: a practical diagnostic tool for future BEL research.

Evaluation should explicitly account for mention-level difficulty.

Key findings

- All models degrade *systematically* on difficult mention types.
- Lexical variation is the dominant challenge, especially for gene/protein abbreviations and rare forms.
- Mention characteristics are proxies for training sparsity.
- Despite contextual encoders, models still rely on string similarity and entity popularity.

Acknowledgements & Contact

Acknowledgements

This work was supported by the **FAIRclinical** project:

- CHIST-ERA grant CHIST-ERA22-ORD-02
- ANR project ANR-23-CHRO-0008-01
- ANR project ANR-22-CPJ1-0087-01

Contact

Baptiste Pras

baptiste.pras@universite-paris-saclay.fr

LISN, Université Paris-Saclay, CNRS

Thank you!

GitHub Repository