Relieving Memory Impaired Named Entity Recognition using Phonological Similarity of Mentions

Avi Bleiweiss
BShalem Research
Sunnyvale, CA, USA
avibleiweiss@bshalem.onmicrosoft.com

Abstract

Consider the brain issuing a request to retrieve a named entity from a neural memory cell. Without deficiency, the memory cooperates to yield what was expected of it— a complete mention sequence. However, for elderly people, in many cases the memory response is distorted and broken, leaving the person suspended. In this paper, we hypothesized that the objective named entity and the corresponding human guess utterance are related phonologically. Our novelty lies in employing entity linking modeling to recover deprived named entity recognition. We utilized a recently introduced high-quality gold-standard dataset for entity linking, and expanded it by transforming textual mentions into phoneme representation. In our evaluation, we used the Levenshtein distance metric in phonological space for measuring similarity between a test subject guess and top-k goal entities. The results we report are compelling to draw the research community attention.

1 Introduction

The formal definition of a named entity in natural language processing (NLP) is often tied to the rigid designator intuition considered a major contribution to the philosophy of language (Kripke, 1980). A rigid designator is a term that refers to the same object in every possible world where that object exists. However, due to high variability the rigid designator concept is unsustainable for memory-impaired phonological named entity retrieval. The task of named entity recognition (NER) further identifies and classifies entities in raw text into predefined categories. In our paper, we chose an NER class schema that is based on OntoNotes 5.0 (Pradhan et al., 2012) which facilitates richly annotated linguistic data.

Entity Linking (EL) is an essential information retrieval task that disambiguates mentions of proper named entities in unstructured text and links them to their corresponding entries in a large knowledge base (KB). An EL dataset is typically annotated with entries that consist of 1) the mention content, 2) the named entity class, and 3) a pointer to an English Wikipedia page. Our work requires in addition a phonological representation of a mention for conducting similarity computation between a pair of utterance phonemes. A sample of our four-element augmented EL dataset entry follows:

mention: `Huckleberry Finn`
class: `PERSON`
target: `Huckleberry Finn`
phon: `HH AH1 K AH0 L B EH2 R IY0 F IH1 N`

We used the ARPABET English pronunciation phoneme set, a transcription system that represents English speech sounds using ASCII characters.

A memory retrieval deficit refers to the temporary inability to access stored information when needed, even though the memory trace may still exist in the brain. Counter to a memory loss of which the data requested is no longer stored, in retrieval deficit the memory is available but inaccessible. In one scenario, named entity recognition is impaired and the initial utterance emitted by an individual as a guiding guess is only remotely related to what was expected as the ultimate desired name. This form of memory deficit misinterprets a person uttering a named entity substitution that is difficult to understand in a broader context. Our study offers a framework to bridge the apparent entity recognition gap by searching top-k named candidates in an EL dataset that are the most phonetically similar to the guess expression. Our similarity computation uses the Levenshtein edit distance and is performed in phonetic space. We note that a human subject with an impaired NER may be well cognizant of the objective entity class which helps reduce the top-k search space.

Our work uses the EL dataset offered by Olewniczak and Szymański (2025), who proposed

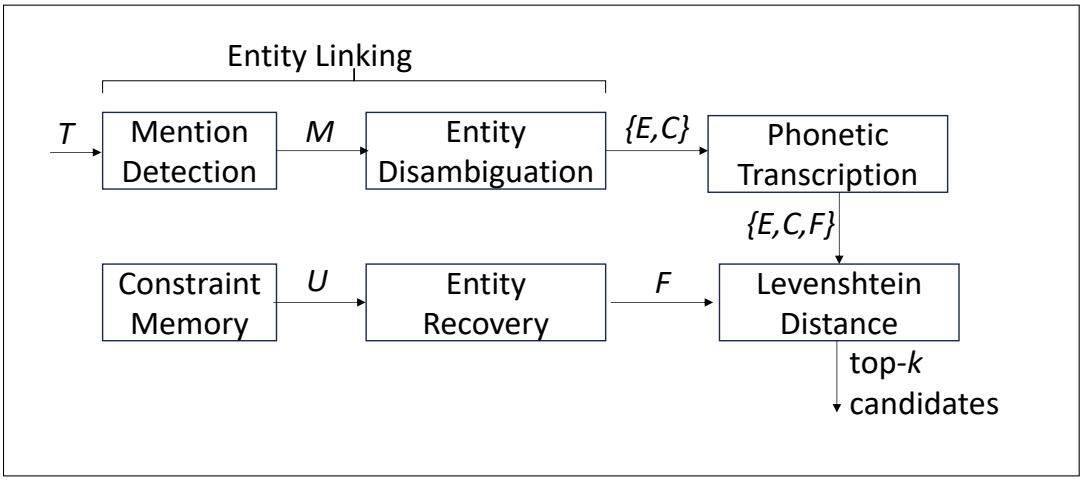

Figure 1: Our framework is founded on the concept of entity linking and augments its scope to support mentions in phonological representation. Formally, the process takes in unstructured text (T) and emits a set of mentions M. Every qualified mention is linked with a named entity E and a context, or class C that resides in the training knowledge base. The second part of the pipeline takes in an utterance U produced by an impaired-memory human with difficulty to properly recover named entities. The utterance is transcripted to textual phonemes F followed by traversing the entity linking dataset and computing the Levenshtein edit distance in phonological space between the utterance and mention phoneme rendition F. The top-k candidates most similar to the utterance are returned to the human subject for conducting named entity matching.

a new gold standard for EL— Elgold. In Elgold, every mention is assigned to a proper class after a rigorous review by humans. Elgold follows the OntoNotes 5.0 class schema with modifications to retain the rigid descriptor abstraction. As such, Elgold avoids classes that link to specific numerical values or points in time that are non-definitive and inaccurate and thus unfit for EL context. The list of Elgold classes is provided in Table 1. Most prevalent EL classes for our study of impaired NER humans are the person and geo-political entities. This aligns with elderly people who often tend to recall names of friends, family members, and renown figures from the past; as well as places they toured that impact their lives. However, to retain generality, especially in scenarios where the objective class is unknown, our framework traverses the entire EL dataset and derives the top mentions that match the guess across all classes of the schema.

To the extent of our knowledge, employing phonological modeling to recover memory obscured named entities has not been explored in prior work. Our contribution includes 1) We extend the EL gold standard dataset to support phoneme representation of mentions. 2) To commence similarity ranking, our task requires that mentions are distinct without repetition across all entity classes. 3) We offer the collection of guess entities we used in our experiments to construct a knowledge base for machine learned goal entities.

2 Resources

In this paper, we used Elgold— a new dataset designed for the evaluation of entity-linking systems (Olewniczak and Szymański, 2025). [1] The dataset contains 276 multi-genre texts with marked named entities that are linked to corresponding English Wikipedia articles. The complete list of English NER classes we used in our paper follows a modified version of the OntoNotes 5.0 schema, [2] and is provided in Table 1 (Pradhan et al., 2012). We show both the raw and unique mention distributions across the 14 EL classes, with a total of 3,559 mentions of which 2,109 are unique. As expected, the person and geo-political classes are the most mention-populated.

To ensure evaluation consistency for conducting top-ranked similarity computation between a guess utterance and class mention, our task requires that mentions we extract from Elgold, or for that matter from any other EL dataset are distinct within each class boundaries. In Table 1, we show Elgold possesses a considerable raw mention decline of almost forty percent, on average, due to replication.

We expand on the entity format we retrieved from Elgold by adding a phonological representation to a textual entity mention. To this extent, we used the CMU Pronouncing Dictionary (CMU-

[1] https://doi.org/10.34808/8zbz-1e66
[2] https://doi.org/10.35111/xmhb-2b84

Class	Code	Raw	Unique
Disease	DISEASE	216	88
Event	EVENT	86	68
Facility	FAC	90	68
Geo-political	GPE	626	279
Language	LANGUAGE	12	6
Law	LAW	22	19
Location	LOC	105	62
Nationality	NORP	124	75
Organization	ORG	551	326
Person	PERSON	748	516
Product	PRODUCT	235	144
Specie	SPECIE	230	132
Substance	SUBSTANCE	310	157
Work of Art	WORK_OF_ART	204	169
TOTAL		3,559	2,109

Table 1: Raw and unique mention distributions across named entity classes with the person class underlined.

dict), a machine-readable pronunciation dictionary that provides orthographic-to-phonetic mappings for English words, using a modified ARPABET phoneme set. In practice, we used cmudict 1.1.3, a versioned Python wrapper package for the CMUdict data files. [3]

3 Method

In Figure 1, we highlight our method for aiding memory impaired entity retrieval. Given input raw text T comprising a set of n recognized entity mentions $m_{1:n}$ and a target KB containing a collection of k named entities $e_{1:k}$, EL aims to map each entity mention m_i to a corresponding target e_j and assign it an NER class c_l based on OntoNotes 5.0 schema definition $c_{1:N}$ with $N = 14$ classes. We extract entity items (m_i, e_j, c_l) from each of the 276 text files provided by Elgold and assemble them into our working set. Our final step transforms each textual mention m_i into a phonetic representation in phonemes f_i and forms our entity item as (m_i, f_i, e_j, c_l). [4] Our working set is further split by the entity class c_l into $N = 14$ item clusters.

The core of our model takes in an utterance u of a named entity guess emitted by a test subject with constraint memory. u is either represented as an oral or written expression that we convert to a machine-searchable pronunciation. In some scenarios, utterance u might be provided with an accommodating named entity class c_l that helps to considerably reduce the search space for phonetically guess-related entity candidates. In this study, we expect the utterance u to be paired with an NER class. Based on this class, we iterate a working set cluster and compute the Levenshtein edit distance between a phonetically mention representation f_i and the transformed guess utterance u.

4 Experiments

Given a guess utterance of a named entity, we calculate top-k phonetically resembling entities to guide a memory-impaired person toward successful retrieval of an obstructed objective entity. Ideally, the test subject identifies one of the top-k entities to match the elusive goal entity. We report our entity similarity results using the Levenshtein distance metric (Levenshtein, 1966). [5] In the context of phonemes, it measures the minimum number of phoneme-level edits— insertions, deletions, and substitutions— to convert one phonological entity representation to another. We note that our model processes any number of phonemes that constitute a named entity mention. We follow with a list of typical goal-guess pairs of named entity mentions.

Goal	Guess
`Wembanyama`	`Alibaba`
`Calgary`	`Gilroy`
`Mollie Stone`	`Monte Python`
`Crate & Barrel`	`Barnes & Noble`

Widely used in NLP, the Levenshtein edit distance is most appealing to our task owing to its fuzzy string matching capability. Due to the two-dimensional matrix nature of the computation, both the time and space complexity of the Levenshtein algorithm are $\mathcal{O}(mn)$, where m and n are the lengths of the two text sequences. Although quantitatively expensive, the relatively short span of mention sequences for which similarity is calculated mitigates the algorithm weakness. Derived from the edit distance, the Levenshtein algorithm also computes a normalized similarity score in a $[0, 1]$ range— the closer to 1 the higher the similarity. In our results, we report both the distance and ratio measures.

We evaluated the performance of our model for each its discovery phase and the recovery step sep-

[3] https://pypi.org/project/cmudict/
[4] https://github.com/bshalem/fon
[5] https://pypi.org/project/Levenshtein/

Mention	Distance	Ratio	ID
Chuck Berry	16	0.67	2102
Michael Cain	16	0.65	2192
Philadelphia Phil	17	0.65	1720
Helen Mirren	14	0.65	2111
Russell Brookes	16	0.64	1975
Ronald Reagan	14	0.63	761
Michael Jackson	15	0.62	2097
Natalie Maines	17	0.62	1678
Pamela Webb	16	0.61	2173
Abigail Breslin	19	0.61	507

(a) Guess: `Huckleberry Finn`; Goal class: PERSON.

Mention	Distance	Ratio	ID
Roman Republic	17	0.62	3444
Waltham Forest	16	0.62	1892
Roman Empire	13	0.62	3448
Singapore	11	0.60	2111
Los Alamitos	15	0.59	1035
New York	12	0.58	993
Poland	12	0.58	3529
Stony Point	12	0.57	2275
California	13	0.56	874
Nepal	13	0.56	1363

(b) Guess: `Rosa Parks`; Goal class: GPE.

Table 2: Phonetically similar top-k named entity candidates extracted from each the PERSON and GPE objective classes. The ratio measure is shown in a descending order, and the ID is an item index of the original EL dataset.

arately. Using a seed intuition to a test subject who utters a guess entity, we search a goal NER class within our post-processed working set, and compute a descending order of similarity ranking in phonological space between the transformed human seed and the mentions of the goal entity class. In Table 2, we show the ten most similar entity candidates to each of the guess mentions `Huckleberry Finn` and `Rosa Parks`. The guess entity and the goal search space In Table 2a are both confined to the person entity class. In another scenario depicted in Table 2b, we show the guess entity is a person and the objective is constricted to the geo-political NER class. The Levenshtein ratio for single class settings averages 0.64 and ranges from 0.61 to 0.67, while crossing an entity class boundary sets a lower similarity mean 0.59 and range $[0.56, 0, 62]$.

In the recovery step, the test subject either identifies the concealed objective entity he expected in the top-k entity list, or not. This task is rather reduced to a simple binary classification.

5 Conclusion

In this paper, we expand on the line of research that addresses memory impaired entity retrieval. We explored the intuition of phonetic resembling to discover memory inaccessible and otherwise obstructed entity mention, The framework we proposed is tied to the entity linking concept and draws mentions from a gold standard dataset. Our results corroborate with our hypothesis for real-world scenarios of temporarily brain-obscured entities. A venue to further explore our work is collecting a large diverse set of guess entities and constructing a knowledge base for recovering machine learned objective entities.

Limitations

We acknowledge several limitations in our paper. The EL gold standard dataset from which we extracted our working set is fairly small with a little over two thousand unique mentions. Additionally, only the GPE, ORG, and PERSON entity classes sustain a reasonable sample size to fit our task. In this study, we had about a dozen test subjects for soliciting guess entities yielding about ten guesses a-day per participant. This group of elderly people in their seventies was adequate to analyze our discovery phase, however, for a large scale and diverse study we would need hundreds of test subjects.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful suggestions and feedback.

References

Saul A. Kripke. 1980. *Naming and Necessity: Lectures given to the princeton university philosophy colloquium.* Princeton University Press, Princeton, USA.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Szymon Olewniczak and Julian Szymański. 2025. Gold standard, multi-genre dataset for named entity recognition and linking. *Scientific Data*, 12(1).

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.