Person name disambiguation in news articles: a hybrid method for enhancing entity resolution in Russia-Ukraine war coverage

Nina Khairova[1,*,†] and Anastasiia Mozghova[2,†]

[1] *Umeå University, 90187 Umeå , Sweden*

[2] *Taras Shevchenko National University of Kyiv, 64/13, Volodymyrska Street, City of Kyiv, Ukraine*

Abstract

Text analytics of frequency, context, and media portrayal of individuals in war reporting provides insights into key figures, biases, and socio-political narratives. However, due to named entity ambiguity, the number of unique individuals mentioned does not always align with the total number of PERSON entities identified in the dataset, which leads to reduced accuracy in the text analysis. To address this challenge and improve the accuracy of individual identification while ensuring a more reliable analysis of the dataset, we applied the Damerau-Levenshtein distance metric and machine learning techniques to identify and consolidate mentions of personal named entities in news coverage of the Russian-Ukrainian war in 2022. As a result, we created a comprehensive personal names dictionary containing 6,414 entries, with each entry grouping name variants that refer to the same individual.

Keywords

Named entity resolution, personal named entity, Damerau-Levenshtein distance, word embedding, Russia-Ukraine war, name dictionary, persons mentioned in news

1. Introduction

Named Entity Resolution (NER) or Entity Resolution (ER) is a fundamental task in Natural Language Processing (NLP) that is essential for enhancing the reliability of text analysis, improving large-scale data processing and information retrieval, and even advancing text generation models. Entity Resolution aims to disambiguate and link named entities (such as individuals, organizations, and locations) to a structured knowledge base, unifying mentions of these entities within unstructured text or creating specific dictionaries or taxonomies. Thus, by providing structured information about key elements in the text and ensuring consistency in entity categorization, ER reduces ambiguity in information extraction and facilitates the execution and analysis of downstream NLP tasks.

However, despite decades of extensive research, named entities' ambiguity continues to present significant challenges due to unresolved complexities in accurately identifying and linking entities across diverse texts. These complexities are further compounded by the intricacies of transliteration and translation rules, particularly when dealing with texts from specialized domains that include named entities from multiple languages.

In our study, we examine the ambiguities of personal named entities in news articles covering events related to the Russian-Ukrainian war, as compiled in the Russian-Ukrainian War (RUWA) dataset [1]. Although first and last names should be unified in English-language news about the current war, variations in spelling and formatting, as well as the use of different aliases and

CLW-2025: Computational Linguistics Workshop at 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025), May 15–16, 2025, Kharkiv, Ukraine

* Corresponding author.

† These authors contributed equally.

nina.khairova@umu.se (N. Khairova); mozgova_@knu.ua (A. Mozghova)

0000-0002-9826-0286 (N. Khairova); 0009-0003-6386-5503 (A. Mozghova)

transliterations, continue to present challenges in maintaining the accuracy of studies that utilize these new articles.

By applying the Damerau-Levenshtein distance metric and machine learning approaches, we aim to identify and consolidate mentions of personal named entities in the dataset, creating a unique and comprehensive record for each individual within the personal names dictionary of news discourse on the Russian-Ukrainian war.

2. Related work

The problem of entity resolution has been studied for over 60 years. Initially, most approaches were based on matching rules, which incorporated considerations of syntactic relations [2], such as string similarity and phonetic patterns, to identify entity correspondences [3]. These early methods primarily relied on deterministic techniques and domain-specific heuristics, focusing on resolving entities within well-defined contexts [4].

Over time, as data complexity and volume have increased, the ER task has come to be considered a unsupervised large-scale clustering problem, where records must be grouped into clusters, each representing a unique entity [5]. However, clustering-based ER methods often require pairwise comparisons between records, leading to quadratic time complexity. This becomes computationally intensive as data volume grows, making scalability a significant concern [6]. Furthermore, unstructured text data, especially short texts like tweets or search queries, often lack sufficient context, leading to sparse and high-dimensional representations. This sparsity complicates the clustering process, as traditional algorithms may struggle to find meaningful patterns without adequate contextual information [7]. The authors [8] also highlight that unsupervised clustering approaches may suffer from inaccuracies due to inefficiencies in the clustering step, where the goal is to discover hidden unique entities.

In turn, supervised learning methods for ER require extensive labeled datasets, where entity mentions are manually linked to their corresponding real-world entities. However, obtaining high-quality annotated data poses significant challenges, as it demands domain expertise, substantial human effort, and scalability [9]. Additionally, supervised ER models trained on one dataset often fail to generalize well across different datasets or domains [10]. To address these limitations, more advanced algorithms have been developed to better handle unstructured text data.

The application of word embedding methods for record representation has improved performance in named entity resolution by capturing semantic relationships between words [11]. This capability enables ER systems to identify different terms referring to the same entity. Moreover, integrating word embeddings into deep learning architectures, such as LSTM-CRF models, has been shown to further enhance ER performance [12].

In recent years, research has focused on utilizing LLMs to identify and merge records referring to the same real-world entity [13, 14]. However, even though LLM-based entity resolution methods can achieve similar performance to deep learning methods trained on large amounts of data, these methods still face several limitations. One major challenge is the dependence of these models on pre-trained data, which can introduce biases or inconsistencies when handling domain-specific entity mentions [15, 16]. This issue is particularly relevant to our study, which focuses on a domain-specific context related to the Russia-Ukraine war. Additionally, the computational cost of LLMs remains a concern, particularly for large-scale entity resolution tasks requiring real-time processing. Moreover, fine-tuning and adapting these models for specific datasets necessitate careful consideration of prompt design [17, 18], data augmentation techniques, and strategies for context-dependent entity mentions [13].

One major challenge in NED is handling out-of-vocabulary (OOV) named entities, particularly personal names, which frequently exhibit spelling variations, transliterations, and abbreviations [19]. Several works have explored hybrid approaches combining character-level embeddings [20] and phonetic similarity algorithms to enhance robustness against OOV entity mentions. However, this challenge remains even for traditional word embedding models such as Word2Vec, FastText, and

GloVe. Word2Vec and GloVe operate at the word level, meaning they fail to generate meaningful representations for unseen or rare name variations. While FastText improves OOV handling by leveraging subword information, it still struggles with name disambiguation due to context sensitivity, particularly in cases where multiple individuals share similar name structures. Additionally, contextual embeddings from transformer-based models such as BERT often rely on external knowledge sources and large-scale corpora, which may not sufficiently capture domain-specific or newly emerging personal names [21]. Thus, effective OOV handling in named entity disambiguation remains an open problem.

3. Motivating data

We consider news articles covering the Russian invasion of Ukraine, published by established media outlets between February 2022 and September 2022, collected in the RUWA (Russian-Ukrainian War) dataset [1, 22]. The dataset comprises 16,545 articles from news websites in Ukraine, Russia, Asia, and the USA. These articles are categorized into nine major, well-known, and information-significant events of the Russian-Ukrainian war, such as the beginning of the war, the Bucha massacre, the sinking of the Russian warship, and others. For each news article in the dataset, information is provided about its content, date of publication, source news website, keywords, a link to the original publication, and the event it covers. Additionally, the vast majority of articles include a title, and some also have subtitles [24].

The dataset was compiled from multiple sources, some of which published identical or near-identical articles. This redundancy, if left unaddressed, could distort statistical analyses and NLP model training by artificially inflating the frequency of certain named entities, leading to biases in entity classification and making entity resolution more challenging. To mitigate these issues, we performed a preprocessing step to identify and remove duplicate entries, ensuring the dataset remains consistent and reliable for downstream tasks. Duplicate identification was based on content similarity, considering two primary forms of repetition: instances where the same article was republished on the same website and cases where it appeared on different websites under a different title. As a result of this preprocessing step, the dataset size was reduced by 75 articles.

In the subsequent step, we identified 18 categories of named entities within the dataset's articles. These entity types align with the predefined categories established in the SpaCy library for Python, which provides accurate entity recognition while maintaining high processing speed. Statistical analysis of the RUWA dataset, as illustrated in Figure 1, reveals that the most frequently occurring named entity types are NORP, which appears in 15,225 articles; GRE, which is present in almost all articles in the dataset; ORG, which occurs in 16,030 articles; DATE, which is found in 13,792 articles; and PERSON, which appears in 13,004 articles.

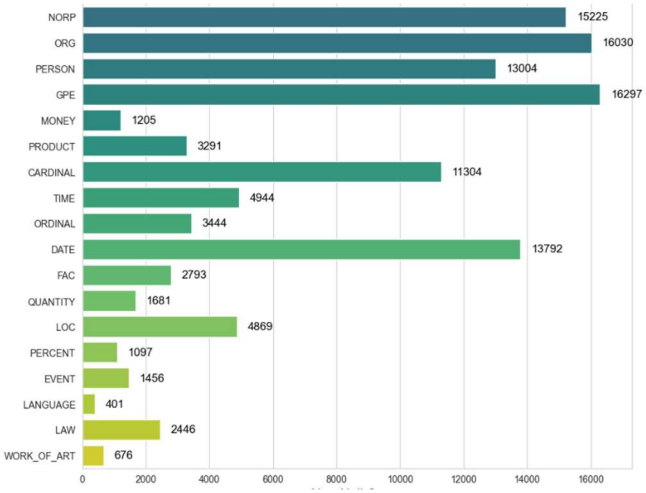

Figure 1: Frequency Distribution of Named Entity Types in the RUWA Dataset Articles.

The NORP category encompasses nationalities, religious groups, and political affiliations, while the GRE category includes geopolitical entities such as countries, cities, and states. The ORG category comprises organizations, including corporations, government agencies, and institutions. The DATE category denotes absolute or relative temporal expressions, and the PERSON category identifies individuals mentioned in the articles.

Building on this analysis, we focus our study on the PERSON category, as it presents distinct challenges compared to other named entity types. Unlike GRE, ORG, DATE, and NORP, which can often be inferred from domain-specific knowledge of geopolitical regions, temporal references, and institutional affiliations, identifying and analyzing individuals in war-related news is considerably more complex. This complexity arises from name ambiguities, variations in spelling, and the use of multiple transliteration schemes from Ukrainian and Russian to English.

In total, we identified 15,131 named entities classified as personal names. Figure 2 illustrates the distribution of the relative frequency of personal name entities per 1,000 tokens across various web news sources. This frequency is calculated for each source by dividing the number of occurrences of personal name entities in that source by the total number of tokens in all articles from that source and multiplying the result by 1,000.

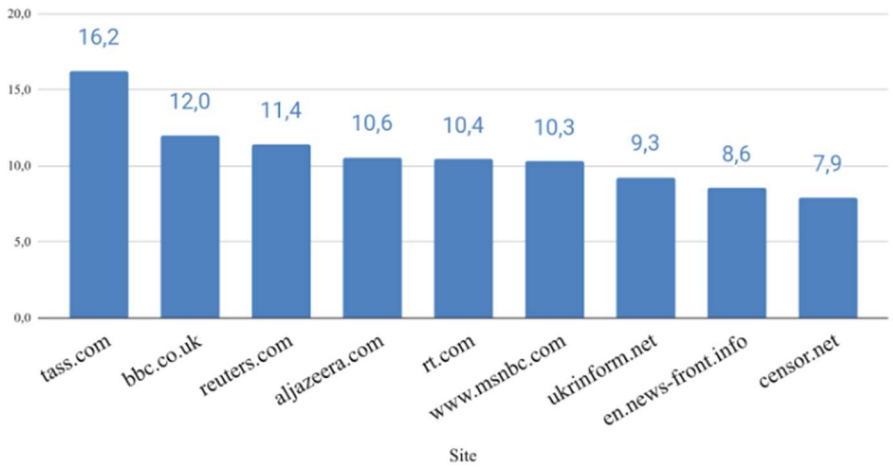

Figure 2: Distribution of the relative frequency of PERSON name entities per 1,000 tokens across various web news sources.

However, the number of unique real-world individuals mentioned in the dataset did not correspond to the total number of PERSON named entities recognized within the dataset. This discrepancy arises from several factors. First, the same individual may be referenced in multiple ways (e.g., by full name, first name only, or initials), resulting in multiple entity instances corresponding to a single real-world person. Second, named entities in the possessive case were sometimes treated as distinct personal entities. Finally, some entities were assigned multiple labels or misclassified under an incorrect entity type. The last two issues were addressed through algorithmic rule-based corrections. Specifically, entities assigned multiple labels simultaneously (e.g., PERSON and GPE or PERSON and ORG) as well as possessive forms were excluded to mitigate errors. Following these corrections, the total number of PERSON-named entities in the analyzed dataset was reduced to 14,574.

As with any string-based similarity measure, edit distance, and word embeddings are highly sensitive to variations in string length and structural differences. These methods rely on the consistency of token structures, and therefore, our analysis was restricted to personal named entities consisting of the same number of tokens. Figure 3 indicates that two-token entities are the most frequent, followed closely by one-token entities, while names consisting of three or more tokens occur much less frequently.

Given that one-token and two-token personal named entities are the most common, we focused exclusively on two-token names. This approach enhances the reliability of entity resolution by reducing the risk of homonymy, which is more prevalent among single-token names. Additionally, it aligns with the structure of name dictionaries, which typically record both given names and surnames to facilitate identification. Including both given names and surnames in two-token entities provides additional contextual information, helping to distinguish individuals more effectively. While three-token named entities may offer even more specific information about a person, their relatively low frequency in the dataset limits their contribution to the overall analysis.

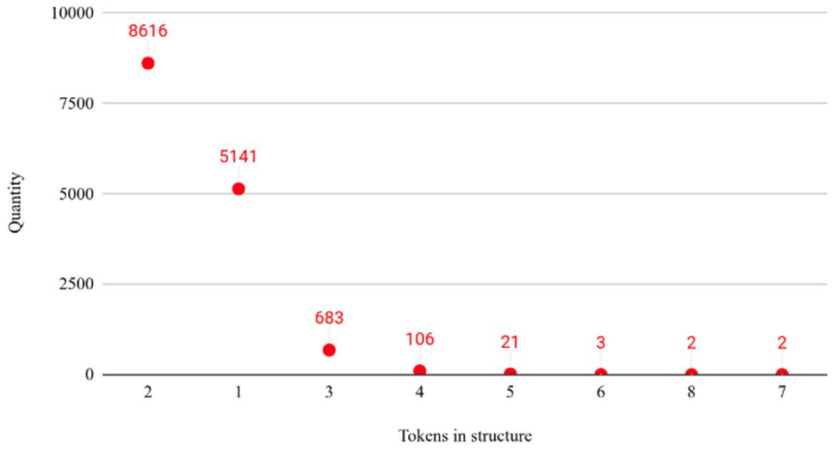

Figure 3: Distribution of personal named entities by the number of tokens in the dataset.

4. Named entity resolution methods

Based on [24], we define the NED task as follows. Given the context of articles in the dataset, where a set of PERSON named entity mentions M has been identified in advance, the objective is to group these mentions into clusters that correspond to unique real-world individuals. Since we did not have access to a structured knowledge base or reference dataset for the individuals mentioned in the war-related articles, the resolution process relied solely on string similarity, contextual embeddings, and co-occurrence patterns to address name variations.

4.1. Damerau-Levenshtein distance

The ambiguity of PERSON named entities in the articles arises not only because the same individual may be referred to by different name variations, such as ['*Joe Biden*', '*Joseph Biden*'], but also due to variations in transliteration from Ukrainian and Russian like *'oleksii reznikov*', *'oleksiy reznikov*', name and surname translation *'olena zelenska*', '*elena zelenskaya'*, as well as typographical errors *'volodmyr zelenskyy'* and *'vladimir zelensy'*.

The problem of different transliterations for the same Ukrainian and Russian names into English leads to confusion in news reports and creates additional challenges for text analysis in the RUWA dataset, which collects news articles from around the world in English but focuses on events involving Russian and Ukrainian personal names. Recent studies have highlighted that these transliteration inconsistencies arise from multiple factors, including linguistic differences, varying transliteration standards, geopolitical influences, and historical conventions [25, 26]. Moreover, the presence of multiple transliteration systems, including ISO 9, ALA-LC, and various national standards, exacerbates inconsistencies in the representation of personal names, causing troubles for both news interpretation and automated text processing.

As an initial step in personal named entity disambiguation, addressing typographical errors, alternative spellings, and minor transliteration inconsistencies, we employed the Damerau-Levenshtein distance (DL distance). In contrast to the well-known Levenshtein distance approach,

which considers only insertions, deletions, and substitutions, the DL distance approach extends the Levenshtein model by also accounting for transpositions and adjacent character swaps [27] that frequently occur in spelling variations, transliteration inconsistencies, and typographical errors in our dataset articles.

We analyzed only two-token personal named entities and computed the DL-distance for each pair of detected two-token PERSON entities. Entities exhibiting a small edit distance were grouped as potential variants of the same individual. Our empirically determined minimal edit distance (≤ 3) is based on the observation that most spelling variations, transliteration inconsistencies, and typographical errors in two-token personal named entities involve a small number of character-level modifications. Based on observations from the RUWA dataset, we determined that a lower threshold (e.g., ≤ 1) was too restrictive and missed many valid name variants, while a higher threshold (e.g., > 3) increased recall but introduced false positives by grouping distinct individuals with the same names, but different short surnames such as *'John Bird'* and *'John Tory'*. For example, different transliteration systems may represent the Ukrainian name *'Volodymyr Zelensky'* as *'Volodymyr Zelenskyi'* or *'Volodymyr Zelenskyiyy,'* depending on the system used. Common errors, such as adjacent character swaps (e.g., "*Baiden*" instead of "*Biden*"), typically involve single transpositions. Generally, these variations consist of minor character insertions, deletions, or substitutions, which can all be quantified within an edit distance of 3 or less.

By computing the DL distances between different name variants, we performed clustering of names exhibiting a small DL distance (≤ 3), suggesting a high probability of referring to the same individual.

4.2. Word embedding approaches

Word embeddings are a powerful tool for representing words as dense vectors in a continuous vector space, where semantically similar words are mapped to nearby points. The similarity between embeddings allows us to check whether two names within the same cluster appear in similar contexts, indicating that they are contextually or semantically related, which facilitates entity resolution. We utilize four different word embedding models: *fasttext-wiki-news-subwords-300*, *word2vec-google-news-300*, *glove-wiki-gigaword-300*, and *glove-twitter-200*.

The FastText model (*wiki-news-subwords-300*) is trained on Wikipedia data and captures subword information, enabling it to effectively handle morphological variations of words. By incorporating subword information, FastText can generate robust embeddings for name variants that may contain typographical errors or less common spellings.

The Word2Vec model (*word2vec-google-news-300*) is trained on a large corpus of Google News data and uses either the Skip-gram or Continuous Bag of Words (CBOW) model to produce embeddings that capture syntactic and semantic relationships between words. The GloVe model (*glove-wiki-gigaword-300*) is trained on a combination of Wikipedia and the Gigaword corpus and generates word embeddings that reflect global word co-occurrence statistics. Both of these models are beneficial for capturing relationships between names and identifying semantic similarities, even in cases of transliteration inconsistencies or different spellings.

Utilizing the *glove-twitter-200* model enables handling informal language and slang, making it useful for resolving name variants in news contexts that involve informal usage or typographical errors, as it is trained on a Twitter corpus.

However, when applying the models mentioned earlier to personal name entity resolution, we encounter the issue of OOV personal names. A significant proportion of the names of individuals from Ukraine and Russia mentioned in news articles on the Russian-Ukrainian war are absent from the training corpora of these models. This can be attributed to several factors. First, pre-trained word embedding models are typically trained on general-purpose corpora, such as Wikipedia, news datasets, or web text, which may not comprehensively cover personal names, particularly those from specific geopolitical or war-related contexts. Second, linguistic and transliteration variations of Slavic names further exacerbate the OOV problem, as different spellings or transliterations may not be

represented in the models' vocabularies. Third, names with low frequency or those associated with lesser-known individuals may not appear frequently enough in the training data to be effectively captured by these embeddings.

Simultaneously, the presence of OOV words, particularly personal names absent from the training data, poses a substantial challenge to the performance of both named entity recognition and named entity resolution, potentially reducing their accuracy and robustness [29]. Since pre-trained models are unable to generate vector representations for certain named entities in the RUWA dataset, we train FastText, Word2Vec, and GloVe on this dataset's texts to enhance their ability to encode domain-specific personal names and mitigate the OOV issue observed in pre-trained embeddings.

To address name variation, we employ a clustering approach that groups different representations of the same personal name based on word embedding similarity. The clustering process is conducted in a vector space, where semantically and contextually similar personal name variants are grouped using cosine similarity, which quantifies their proximity in the embedding space. This method allows us to capture morphological, phonetic, and transliteration-induced variations while preserving contextual relationships among names.

The resulting clusters facilitate the disambiguation and consolidation of personal name mentions across news texts, ultimately forming a name dictionary that aggregates all detected name variants around the most frequently mentioned form. This dictionary serves as a reference resource for standardizing named entities in news coverage of the Russian-Ukrainian war.

5. Experiments

In this section, we conduct thorough experiments to investigate two strategies for enhancing personal named entity resolution. First, we explore the use of the DL distance to mitigate ambiguity in personal named entities caused by typographical errors, alternative spellings, and minor transliteration inconsistencies (§ 4.1). Then, in § 4.2, we explore the application of a word embedding approach that utilizes distributed representations to capture semantic relationships among personal names in the RUWA dataset.

5.1. Leveraging the DL distance metric

This step takes as input a collection of personal named entities and applies the DL distance to measure the similarity between different name variants. By grouping names that exhibit minor variations, such as typographical errors, alternative spellings, or transliteration inconsistencies, we created a name dictionary of individuals mentioned in news articles on the Russian-Ukrainian war.

To evaluate the effectiveness of disambiguation using the Damerau-Levenshtein distance metric, we randomly sampled 500 name variants from the constructed Name Dictionary. Two domain experts then manually annotated these variants to create a gold standard reference dataset, ensuring accurate identification of name variant clusters. During this process, 539 unique individuals were identified, as some name variants had been incorrectly grouped and required reallocation into distinct additional clusters.

To evaluate the effectiveness of the DL distance metric for named entity resolution, we use the macro-averaged F1-score. In the standard formulation of the F1-score [30], we define the number of correctly clustered personal named entities in the evaluation matrix as true positives (TP), the number of missing named entities as false negatives (FN), and the number of incorrectly clustered personal named entities as false positives (FP). Since named entity resolution involves multiple entity classes with imbalanced distributions, we adopt the macro-averaging approach, which calculates the F1-score for each entity cluster independently and then averages the scores across all clusters. This method ensures that each class contributes equally to the final score, preventing dominance by high-frequency entities and providing a more balanced evaluation of performance.

Evaluation of the DL distance in measuring the similarity between different name variants and grouping them into clusters results in a macro-averaged F1-score of 0.96. The F1 score, as the harmonic mean of precision and recall for each cluster, provides a balanced measure of the model's

performance, reflecting its ability to maintain both high precision and completeness in the entity resolution process.

5.2. Leveraging word embeddings

Although effective, the edit distance approach does not account for semantic relationships. To address this limitation, we employ embedding models, which represent named entities as dense vectors in a high-dimensional space, where semantically similar entities are mapped to nearby locations. The embedding models capture contextual meanings and semantic nuances that are crucial for disambiguating names.

For the experiment, we used cosine similarity to measure the similarity between name variants represented as word embeddings. By clustering name variants based on this similarity, we aimed to group semantically related names, even if they did not share identical spellings or linguistic forms.

However, before proceeding with the main experiment, we first examined the impact of the OOV issue on the task of personal named entity resolution within the RUWA dataset. To investigate this, we randomly selected 241 PERSON-named entities from a list of individuals mentioned in news articles about the Russian-Ukrainian war. These names were then vectorized consistently across all four models to ensure a comparative evaluation of their representation capabilities. The bar chart in Figure 4 illustrates the relationship between the number of OOV and in-vocabulary one-token personal named entities across the various models.

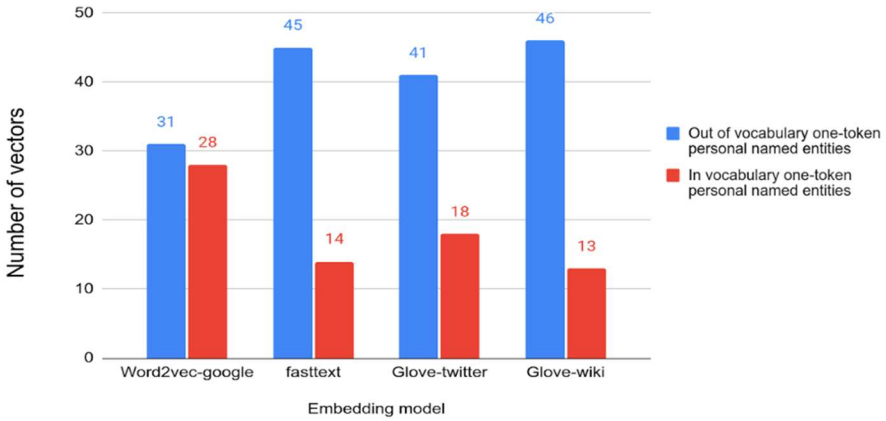

Figure 4: The distribution of out-of-vocabulary and in-vocabulary one-token personal named entities across four different models.

Our experiment revealed that for one-token personal names, which represent individual name components (e.g., first names or surnames) in news articles, the coverage of personal named entities by the Word2Vec-Google, FastText, GloVe-Twitter, and GloVe-Wiki models varies from 47.5% (for the Word2Vec-Google model) to 22.0% (for the GloVe-Wiki model). This suggests that while names or surnames may be present in the training data of the FastText, Word2Vec, and GloVe models, their coverage remains highly limited. Furthermore, all 182 two-token personal name entities that we randomly extracted were classified as OOV by all four predefined models.

To mitigate the OOV issue, we trained the FastText and Word2Vec models specifically on the RUWA dataset. The training corpus consisted of all articles, descriptions, titles, and summaries derived from RUWA, following the necessary preprocessing steps. The word embedding models were then trained using the Continuous Bag of Words (CBOW) architecture, which was selected for its efficiency in capturing contextual relationships.

To ensure consistency and enhance performance in capturing relationships between words, a dimensionality of 200 for the trained vectors was applied to each of the models. For FastText and Word2Vec, we set the training to 5 epochs and employed the CBOW approach, based on the hypothesis that it would optimize performance. For GloVe, we used 50 epochs and a window size of

15. As a result of the experiment, we obtained six distinct embedding models, which included combinations of FastText, Word2Vec, and GloVe models trained with various configurations.

Based on the best performance of FastText in the evaluation, we utilized the FastText model for vector representation to cluster personal named entities based on cosine similarity. Empirically, a cosine similarity threshold of 0.85 was established to identify personal name variants as belonging to the same individual. This threshold was chosen based on its ability to balance precision and recall in grouping semantically similar name variants while minimizing the risk of overgeneralization.

As a result, we obtained 5,117 clusters of personal names. The largest cluster contains 94 personal names, while the smallest cluster consists of a single personal name. Notably, 80% of all clusters include only one personal name. The mean number of name variants per cluster is 1.5901, which is marginally higher than the clusters obtained using the edit distance approach. As a result of the evaluation described in Section 5.1 for the DL distance metric, we achieved a macro-averaged F1 score of 0.91 for the FastText model trained on the RUWA dataset.

5.3. Results

Disambiguation of personal name entities through the application of the DL distance metric and word embedding approach facilitated the creation of the open-recourse dictionary of personal names in news discourse related to the Russian-Ukrainian war [31]. The dictionary comprises 6,414 entries, each representing a set of name variants likely referring to the same individual. The name dictionary is organized as a two-column table, with the first column containing the most frequent variant of each personal name, designated as the canonical form, and the second column listing the corresponding name variants. This structured format supports efficient entity resolution by consolidating different spellings and transliterations of the same individual.

The largest dictionary entry, containing 22 variants associated with a single individual, pertains to different forms of the personal name Volodymyr Zelensky. Despite this cluster's dominance in terms of the number of variants, it is noteworthy that the majority of clusters consist of only a single variant, which collectively accounts for 83% of the entire dictionary. Consequently, the mean number of name variants per individual in the RUWA dataset is approximately 1.3. Furthermore, a clear inverse correlation is observed between the number of name variants for a given individual and the frequency of those variants: as the number of variants increases, the number of individuals associated with those variants decreases within the dictionary.

Table 1 presents a fragment of the personal names dictionary from the news discourse on the Russian-Ukrainian war, highlighting the nine most frequently mentioned individuals in the news articles of the RUWA dataset.

6. Discussion and conclusion

This study addresses the challenge of personal named entity ambiguity within the Russian-Ukrainian war-related news discourse, employing a hybrid approach that integrates the Damerau-Levenshtein distance metric with machine learning techniques. An evaluation of both approaches, conducted using a manually curated set of 542 groups of personal name variants, demonstrated that the edit-distance approach outperformed the word-embedding model in terms of F1 score.

Our methodology facilitated the creation of an open-resource structured dictionary of personal names, comprising 6,414 entries, which effectively consolidates multiple name variants referring to the same individuals. This resource enhances entity resolution within the RUWA dataset, thereby improving the accuracy of analyses related to individual frequency and media representation. Further analysis of these frequencies, along with the context and media portrayal of individuals, will provide valuable insights into the representation of key figures in war reporting, the identification of potential biases, and the broader socio-political narratives surrounding key figures in the conflict.

Table 1
Fragment of the personal names dictionary in news discourse on the Russian-Ukrainian war

Most Frequent Personal Name	Variants
vladimir putin	['vladimir putin', 'vladimir lenin', 'vladimir potanin', 'vlamidir putin', 'vladmir putin', 'valdimir putin']
volodymyr zelensky	['volodymyr zelensky', 'volodymyr zelenskyy', 'volodymyr zelenskyi', 'volodymyr zelenskiy', 'volodmyr zelenskyy', 'volodymr zelenskiy', 'volodymyr zelenksiy', 'volodymyr zelens'kyi', 'volodymyr zelenski', 'volodymyr zelenskyiyy', 'volodymyr zelenskuy', 'volodymyr medinsky', 'voldymyr zelenskiy', 'volodomyr zelenskiy', 'volodymir zelensky', 'voldymyr zelensky', 'volodomyr zelensky', 'volodymr zelenskyy', 'volodomy zelenskyy', 'volodomyr zelenskyy', 'volodynyr zelenskiy', 'volodomir zelenskyy']
joe biden	['joe biden', 'james biden', 'joseph biden', 'jill biden', 'joe bidening']
dmytro kuleba	['dmytro kuleba', 'dmitry kuleba']
pavlo kyrylenko	['pavlo kyrylenko', 'pavlo kirilenko', 'pavel kyrylenko', 'pavlo kyrylenko']
antony blinken	['antony blinken', 'anthony blinken', 'tony blinken']
boris johnson	['antony blinken', 'anthony blinken', 'tony blinken']
emmanuel macron	['emmanuel macron', 'emanuel macron']
iryna vereshchuk	['iryna vereshchuk', 'irina vereschuk', 'irina vereshchuk', 'iryna vereschuk', 'iryna vershchuk', 'iryna vereshschuk']

The findings also indicated that pre-trained word-embedding models struggle to generate accurate vectors for personal names consisting of multiple tokens and frequently fail to generate vectors for single-token entities as well. Furthermore, the study highlights the inherent challenges associated with personal name disambiguation, particularly in the context of transliterations and variant forms across languages. These challenges underscore the necessity for advanced techniques in disambiguation to ensure the accuracy and reliability of named entity resolution in complex datasets..

Declaration on Generative AI

During the preparation of this work, the authors used X-GPT-4 in order to: Grammar and spelling check. After using service, the authors reviewed and edited the content as needed and takefull responsibility for the publication's content.

References

[1] N. Khairova, B. Ivasiuk, F. L. Scudo, C. Comito, and A. Galassi, A first attempt to detect misinformation in Russia-Ukraine war news through text similarity. Proceedings of the 4th Conference on Language, Data and Knowledge (LDK), 2023. Pp. 559–564.

[2] J. R. Finkel, C. D. Manning, Joint parsing and named entity recognition. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics 2009 Jun, pp. 326-334.

[3] C. Dozier, R. Kondadadi, M. Light, A. Vachher, S. Veeramachaneni, R. Wudali, Named entity recognition and resolution in legal text, Springer Berlin Heidelberg, 2010, pp. 27-43.

[4] P. Kalamkar, A. Agarwal, A. Tiwari, S. Gupta, S. Karn, V. Raghavan, Named entity recognition in indian court judgments. arXiv preprint arXiv:2211.03442. 2022 Nov 7.

[5] S. Tipirneni, R. Adkathimar, N. Choudhary, G. Hiranandani, R. A. Amjad, V. N. Ioannidis, C. Yuan, C. K. Reddy, Context-Aware Clustering using Large Language Models. arXiv preprint arXiv:2405.00988, 2024 May 2.
[6] V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, K. Stefanidis, An overview of end-to-end entity resolution for big data. ACM Computing Surveys (CSUR), 2020 Dec 6;53(6):1-42.
[7] M. H. Ahmed, S. Tiun, N. Omar, N. S. Sani, Short text clustering algorithms, application and challenges: A survey. Applied Sciences, 2022 Dec 27;13(1):342.
[8] I. Akef Ebeid, J. R. Talburt, A. S. Siddique, Graph-based hierarchical record clustering for unsupervised entity resolution. arXiv e-prints., 2021 Dec:arXiv-2112.
[9] J. Wu, A. Sefid, A. C. Ge, C. L. Giles, A supervised learning approach to entity matching between scholarly big datasets. In Proceedings of the 9th Knowledge Capture Conference, 2017 Dec 4, pp. 1-4.
[10] R. Wu, S. Chaba, S. Sawlani, X. Chu, S. Thirumuruganathan, ZeroER: Entity resolution using zero labeled examples. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2020 Jun 11, pp. 1149-1164.
[11] N. Kooli, R. Allesiardo, E. Pigneul, Deep learning based approach for entity resolution in databases. In Asian conference on intelligent information and database systems, Springer International Publishing, 2018 Feb 14, pp. 3-12.
[12] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, U. Leser, Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics, 2017 Jul 15;33(14):i37-48.
[13] H. Li, L. Feng, S. Li, F. Hao, C. J. Zhang, Y. Song, On leveraging large language models for enhancing entity resolution. arXiv preprint arXiv:2401.03426. 2024 Jan 7.
[14] T. Wang, X. Chen, H. Lin, X. Chen, X. Han, H. Wang, Z. Zeng, L. Sun, Match, Compare, or Select? An Investigation of Large Language Models for Entity Matching. arXiv preprint arXiv:2405.16884, 2024.
[15] F .Wang, W. Mo, Y. Wang, W. Zhou, M. Chen, A causal view of entity bias in (large) language models. arXiv preprint arXiv:2305.14695, 2023 May 24.
[16] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, Ahmed NK. Bias and fairness in large language models: A survey. Computational Linguistics, 2024 Jun 11:1-79.
[17] N. Nananukul, K. Sisaengsuwanchai, M. Kejriwal, Cost-efficient prompt engineering for unsupervised entity resolution in the product matching domain. Discover Artificial Intelligence. 2024 Aug 16;4(1):56.
[18] Y. Xia, J. Chen, X. Li, J.Gao, APrompt4EM: Augmented Prompt Tuning for Generalized Entity Matching. arXiv preprint arXiv:2405.04820, 2024 May 8.
[19] P. McNamee, H. T. Dang, Overview of the TAC 2009 knowledge base population track. In *Text analysis conference (TAC)* 2009, November. Vol. 17, pp. 111-113.
[20] M. T. Luong, C. D. Manning, Achieving open vocabulary neural machine translation with hybrid word-character models. 2016, *arXiv preprint arXiv:1604.00788*.
[21] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, Gpt-ner: Named entity recognition via large language models. 2023, *arXiv preprint arXiv:2304.10428*.
[22] N. Khairova, A. Galassi, F. L. Scudo, B. Ivasiuk, F. L. Scudo, I. Redozub, Unsupervised approach for misinformation detection in Russia-Ukraine war news. Ceur workshop proceedings, 2024, v. 3722, pp. 21-36
[23] N Khairova, B Ivasiuk, F LO SCUDO, A Galassi, RUWA: Russian-Ukrainian War Dataset, 2023, GitHub. https://github.com/ninakhairova/dataset_RUWA
[24] A. Moro, A. Raganato, R. Navigli, Entity linking meets word sense disambiguation: a unified approach. Transactions of the Association for Computational Linguistics, 2024, 2, pp. 231-244.
[25] M. Vakulenko, Reversible transliteration of the historical Ukrainian alphabets in the context of heritage preservation and linguistic technologies development. *Qeios*, 2023.
[26] N. Knoblock, Misha or Mihailik: A Sociolinguistic View on the Ukrainization of Russian Proper Names in Modern Ukraine. *Names*, *67*(3), 2019, pp.136-152.

[27] A. Çelebi, A.Özgür, Cluster-based mention typing for named entity disambiguation. *Natural Language Engineering*, 2022, *28*(1), pp.1-37.
[28] A. Çelebi, A.Özgür, Cluster-based mention typing for named entity disambiguation. *Natural Language Engineering*, 2022, *28*(1), pp.1-37.
[29] I. Korostelev, K. Aghakasiri, Named Entity Recognition Performance on Out of Vocabulary Words, 2020. *arXiv preprint arXiv:2005.07628*.
[30] URL: F1 score https://en.wikipedia.org/wiki/F1_score, accessed on 01/03/25
[31] A. Mozghova, RUWA-PersonalNameDict. A dictionary of personal names in news discourse on the Russian-Ukrainian war, 2025. GitHub: https://github.com/stubborn-dog/RUWA-PersonlNameDict