Is Literal Annotation Enough? Building an Annotation Framework for Metonymic Named Entities in Marathi

Pratibha Dongare
The English and Foreign Languages University
pratibhaphdlandp22@efluniversity.ac.in

Abstract

Named Entity Recognition (NER) has been a core task of natural language processing (NLP) since the Message Understanding Conferences (MUCs). Data annotation plays a crucial role in this task. However, existing annotation studies often rely on the literal sense of entities. Such annotations may lead to inconsistencies, while resolving ambiguity introduced by figurative tropes like metonymy. For example, in *India won the series*, *India* refers to a sports team instead of a geographic location. Understanding such non-literal senses is crucial for various NLP applications such as Question Answering, Information Extraction, etc. By addressing this gap, this study presents an annotation framework and detailed guidelines for annotating metonymic readings of named entities in Marathi, an Indo-Aryan language spoken in the central-western region of India. The study uses news corpus from various domains. It presents a two-tiered annotation framework for annotating conventional metonymies in Marathi language. Further, it describes the annotation framework applied to a corpus of 1,279 Marathi sentences. The result shows the inadequacy of literal-only annotation as 53.6% of named entity spans have metonymic readings. This study makes a crucial contribution for resource development for low-resource languages that share similar linguistic structures and cultural contexts. The paper describes the framework with necessary examples, challenges and concludes with a future scope.

Keywords: metonymy detection, named entities, Marathi language

1. Introduction

Named Entity Recognition (NER) has been a core task in NLP since the Message Understanding Conferences (MUC) (Grisham and Sundheim, 1996; Nadeau and Sekine, 2007) and the CoNLL shared task (Tjong Kim Sang and De Meulder, 2003) . The primary objective of NER is to identify and classify proper nouns that refer to real-world entities. Conventional categories include Person (PER), Location (LOC), and Organization (ORG). However, annotations often rely on the literal sense of entities, which can lead to inconsistencies when annotators attempt to resolve ambiguity introduced by figurative tropes such as metonymy. Metonymy is a cognitive and linguistic phenomenon, a figurative trope in one entity is used to refer to another entity associated with it (Johnson and Lakoff, 1980). For instance, in example (1), ***India***, a geographic location, refers metonymically to a sports team.

(1) ***India*** *won the series.*

In another example (2) (McShane and Nirenburg, 2021), ***the spiky hair*** refers to a particular person having spiky hair.

(2) ***The spiky hair*** *just smiled at me.*

Several studies have focused on the metonymy resolution task (Markert and Nissim, 2002a,b; Markert and Hahn, 2002; Markert and Nissim, 2007, 2009; Gritta et al., 2017; Gritta, 2019). The majority of this foundational work focused on English, with a limited attention given to German (Markert and Hahn, 2002) and French (Poibeau, 2006). While annotating named entities, metonymic senses of entities need to be annotated since such instances frequently occur in the text. Markert & Hahn (Markert and Hahn, 2002) noted 17% of metonymic instances in German magazines. This makes the nature of metonymy regular, prevalent, and productive (Markert and Nissim, 2002b).

The present study considers Marathi, an Indo-Aryan language spoken in the central-western region of India. Although several studies have focused on the foundational task of NER for the Marathi language (Patil et al., 2016, 2020; Litake et al., 2022, 2023), specific metonymic readings of named entities are rarely explored. This study addresses this critical gap by presenting annotation guidelines currently being applied to construct a comprehensive framework for metonymy detection in Marathi.

2. Dataset

For the analysis, 1,279 sentences from Marathi news across politics, finance, sports, and travel domains, entirely in Devanagari script were used. The corpus follows the structure of the SemEval 2007 shared task dataset (Markert and Nissim, 2007) and expands the scope of metonymic entity types to include PER and MISC categories in addition to LOC and ORG (Nissim and Markert, 2003; Gritta et al., 2017). Existing Marathi NLP resources such as L3Cube-MahaBERT and MahaCorpus (Litake et al., 2023, 2022) offer strong general-purpose representations for Marathi text. However, these

Proceedings of WILDRE-8, pages 33–38

resources were not adopted for the present corpus. Three key reasons motivate this decision. First, existing Marathi NER datasets follow conventional literal annotation schemes. They do not account for metonymic readings of named entities. Second, this study requires a controlled, domain-specific corpus from contemporary news text. Such a corpus better captures the metonymic patterns relevant to information extraction tasks. Third, the tagset used here extends standard NER categories with an explicit Literal/Metonymic distinction. This distinction is absent from all existing Marathi corpora. The annotated data contains 1,741 named entity spans (2,363 NE tokens). Of these, 933 spans (53.6%) are metonymic and 808 spans (46.4%) are literal. Table 1 presents the span-wise distribution across the four NE categories. Fig. 1 shows the span count by entity type. At the sentence level, 418 sentences (32.7%) contain no named entities. 292 (22.8%) contain only literal NEs, 320 (25.0%) contain only metonymic NEs, and 249 sentences (19.5%) contain both literal and metonymic NEs within the same sentence. This shows that the two readings co-occur frequently in Marathi text. Fig. 2 presents the sentence-level distribution of these senses.

NE Type	Literal	Metonymic	Total
LOC	212	580	792
MISC	79	20	99
ORG	23	313	336
PER	494	20	514
Total	**808**	**933**	**1741**

Table 1: Span-wise distribution of entities.

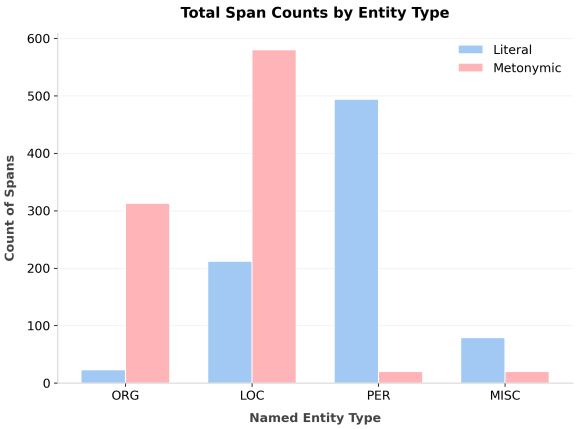

Figure 1: Total span count by entity type

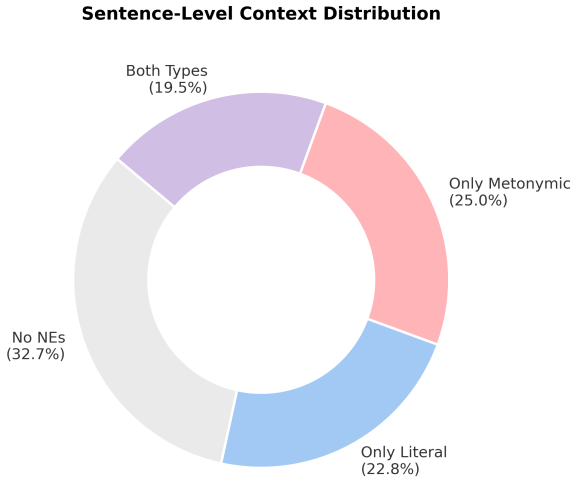

Figure 2: Sentence level distribution of senses

3. Annotation Guidelines and Framework

To achieve high consistency and reliability in metonymy annotation, a principled framework is necessary. Among the previous studies, (Markert and Nissim, 2002b) provided a foundational framework for metonymy annotation in English. This study presents a two-tiered annotation system: first, a set of general principles and second, a specific framework for metonymic interpretations across categories.

3.1. General Principles

These general principles ensure precision in corpus creation, minimizing subjectivity and ensuring high Inter-Annotator Agreement (IAA).

- **Consistency and Uniformity:** *Annotate similar patterns in a similar way across all domains and categories.*
  Apply the same annotation rule uniformly regardless of the position of the entity in the sentence or domain of the text. For example, in (1), if *India* (LOC) referring to its sports team is annotated as a metonymic case of LOC in a sports article, then other entities referring to its team must also be annotated as LOC-metonymy in a different article. If abbreviated forms (e.g., *ISRO*, *UP*) are annotated, then other acronyms (e.g., *UNESCO*, *NATO*) must be annotated using the same logic.

- **Economy:** *Prioritize practicality and efficiency.*
  Avoid fine grained or overlapping labels that increase cognitive load on the annotator(s) resulting in inconsistencies. For instance, this principle is the justification for the No Nested

Entities rule, as a coarse-grained annotation structure is more practical for both annotators and computational models. For example, in (1), *India* is marked as LOC with a metonymic sense and not for the pattern of metonymy (Location-for-organization).

- **Minimal Span:** *Select the smallest possible sequence of tokens that fully identifies the named entity.*
  In a span, include all essential components of the proper name, but exclude contextual modifiers. For example, in the phrase *the former Prime Minister Indira Gandhi*, the minimal span is *Indira Gandhi*. The modifiers *the former*, and the title *Prime Minister*, are all excluded as they are not part of the proper name itself (ref to section 3.2.1).

- **Single Span:** *An entity mention must be annotated as a single, continuous unit*.
  Only the continuous text spans are annotated. Exclude split mentions or discontinuous entities. Split mentions are annotated as separate entities. Example *Students of Mumbai and Pune University* requires two tags. *Pune University* is tagged as an organization, while *Mumbai* is tagged as location metonymy as it is a split mention.

- **Nested Entities:** *Only the outermost, most salient entity in a nested construction is annotated.*
  When an entity is embedded within a larger entity (e.g., *Maharashtra* within *Government of Maharashtra*), annotate only the outermost entity. This prevents ambiguous, overlapping spans and directly supports the Economy and Single Span principles.

- **Default to Literal:** *When in doubt, always annotate the entity as literal.*
  If an entity's usage is ambiguous or could plausibly be interpreted as either literal or metonymic, the entity is tagged as literal. This ensures that the metonymic instances in the final corpus represent high-confidence. For example,

  (3) The name ***Tata*** *is enough.*

  Here, ***Tata*** is ambiguous. It could refer to the person, the organization, or metonymically to the reputation associated with the name. The context alone is insufficient to determine the correct reading. Following the Default to Literal principle, it is tagged as literal PER or ORG depending on the contextual judgment.

- **No Metaphors and Focus on Conventional Metonymy:** *The framework's scope is strictly limited to metonymy within Named Entities.*
  This study targets metonymic sense in named entities (PER, LOC, ORG, MISC). The framework explicitly excludes all other figurative language, such as metaphor. For example,

  (4) *He is the* ***Sachin Tendulkar*** *of his team.*

  The entity ***Sachin Tendulkar*** could mean a metaphor for a great batsman. Non-NE metonymy (unconventional metonymy), such as in example (2) (a Part-for-Whole metonymy using a common noun) is excluded from the annotation.

- **Script & Form Neutrality:** *Annotation decisions are independent of the script, spelling, or form of the entity.*
  This principle helps in annotating a mixed-script (Devanagari/Roman) corpus. For example, the Romanized and Devanagari of acronym *ISRO* (Indian Space Research Organization) are both annotated with the same label (e.g., ORG), as they refer to the identical real-world entity.

- **Data Handling and Annotation Platform:** *Raw text data is annotated using the INCEpTION platform to ensure technical consistency and applicability.*
  The raw, unprocessed news text provided in *.txt* format. To capture the raw, real-world data and linguistic variations, this approach is used. For annotation, INCEpTION (Klie et al., 2018), an open-source, multi-layer annotation platform is used. The platform is user-friendly, offers layer customization, supports for diverse languages and scripts, including Devanagari. This technical choice improves the interoperability and future applicability of the corpus. Fig. 3 shows an example of annotated sentence using this platform. The sentence translates to: *Various international agreements have been criticized for being unfair to the United States.* Here, the *United States* is annotated as metonymic entity.

Figure 3: Example of annotated sentence using INCEpTION platform.

3.2. Category Specific Annotation Scheme

The present annotation framework adopts standard Named Entity categories, accounting for metonymy

across Location, Organization, Person, and Miscellaneous entities.

3.2.1. Entity Span Guidelines

These guidelines refine the general principles to handle complex cases where modifiers are present.

- **Pre/Post Modifiers (Trigger Words):** Modifiers are generally excluded from the entity span as per the Minimal Span principle (e.g., *the*, *former*). However, there are two key exceptions:
  1. **Category-changing Names:** If a modifier combines with an entity to form a new, distinct entity of a different category, the entire phrase is annotated as the new category. For example, in *Temples of Bharat*, *Bharat* is a LOC, but the entire phrase is the proper name of a book (a MISC entity). Similarly, *Shanghai summit* is annotated as a single MISC (event) entity, not as a LOC-metonymy for *Shanghai.*
  2. **Organizational Proper Names:** When a person's name functions as a standard part of an organization entity (e.g., *Modi government*, *Trump administration)*, the entire, continuous span (*Modi government*) is annotated as a single entity.
- **Ambiguity Resolution:** All instances of high ambiguity where the context is insufficient for a clear decision are resolved by applying the Default to Literal principle.

3.2.2. Specific Metonymic Interpretations

The following are the representative examples of category-wise metonymic readings.

- **Location Names:** A location entity is used to refer to an associated institution, an event, or its inhabitants. It includes place-for-government /team. For example,

  (5) ***India*** *hosted the G-20 conference.*

  Here, ***India*** (a location) metonymically refers to the Indian government or organizing body (an organization).
- **Organization Names:** An organization entity is used to refer to its associated personnel, product, or actions. For instance,

  (6) ***France 24*** *posted a video about the event.*

  France 24 (an organization) refers to the employees (a person or a group) of that organization.
- **Personal Names:** A personal entity is used to refer to their associated role, office, or a creation (work, brand, etc.). For example,

  (7) ***Armani*** *still owns the show.*

  Armani (a person) refers to the fashion brand (organization) created by that person.
- **Miscellaneous (MISC):** This category includes metonymic uses of proper nouns that do not fit other three categories. This includes awards, legislation, acts, etc. A common case is an award named after a person, where the entity refers to the object, not the person. For example, (8) *He received the* ***Dadasaheb Phalke Award***. The entity ***Dadasaheb Phalke Award*** is annotated as MISC. It refers to the award itself, which is named after the person *Dadasaheb Phalke*.

The distribution across NE categories reveals striking asymmetries that highlight the need for metonymy annotation. LOC entities are metonymic in 73.2% of spans (580 of 792), and ORG entities in 93.2% of spans (313 of 336). This suggests that a literal-only pipeline would mislabel the large majority of these entities in Marathi news text. PER entities, by contrast, are predominantly literal (494 of 514 spans, 96.1%). Fig. 4 shows the metonymy vs literal proportion per entity class.

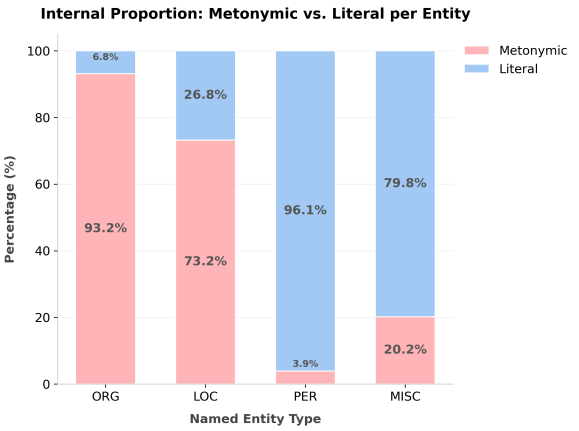

Figure 4: Internal proportion of metonymy vs. literal sense per NE

4. Challenges

The annotation process, even when guided by explicit guidelines, is challenging and complex. This study encountered several challenges, primarily related to linguistic ambiguity, resource constraints specific to Marathi.

4.1. Ambiguity

One of the main challenges is resolving the inherent ambiguity of metonymic usage. While the framework employs the Default to Literal principle to handle confusing cases, contexts often arise where

the metonymic shift is subtle. For instance, in a political domain, deciding whether a location entity like *Delhi* refers to the literal place, the government, or something else requires significant contextual understanding. Subjectivity can lead to inconsistencies and reduced Inter-Annotator Agreement. The Default to Literal principle provides a deterministic resolution strategy in such cases, ensuring that only high-confidence metonymic instances are retained in the final corpus.

4.2. Overlapping of Principles

Creating an effective framework requires balancing detailed instruction with the principle of Economy. Defining too many principles or overlapping guidelines can confuse annotators, leading to errors. For example, the precise interaction between the Minimal Span principle and the handling of pre or post modifiers (trigger words) must be defined carefully. A complex framework, even if theoretically sound, is practically difficult to deploy. This tension was resolved iteratively during the annotation process, with the Economy principle taking precedence when guidelines conflicted.

4.3. Resource and Data Constraints

Working with raw, unprocessed news data in Marathi presents several difficulties. The data is often subject to:
1. **Script and Orthographic Variation:** Spelling variations in the use of script, particularly in the representation of names and acronyms. Such variations and errors affect the annotation process.
2. **Unprocessed Text:** While standard preprocessing usually removes punctuation, this corpus retains the raw text. This results in merged tokens (e.g., punctuation attached to words), which creates significant orthographic noise and complicates the identification of precise entity boundaries. the presence of numerous punctuations and symbols creates noise and complicates the task. The Script and Form Neutrality principle directly addresses orthographic variation. This ensures that annotation decisions remain consistent regardless of spelling or script differences in the raw data.

5. Conclusion and Future Scope

This study presented a principled two-tiered annotation framework for identifying metonymic named entities in Marathi, along with a corpus of 1,279 annotated sentences. The framework establishes eight general principles and category-specific guidelines covering LOC, ORG, PER, and MISC entities. Applied to a Marathi news corpus, the framework reveals that 53.6% of named entity spans are metonymic, with particularly high metonymic rates in ORG (93.2%) and LOC (73.2%) categories. To the best of our knowledge, no prior work has explored an annotation scheme and corpus for metonymy detection in Marathi. The framework is designed to be extensible to other Indo-Aryan languages with similar linguistic structures. Future work will involve completing corpus annotation and training baseline models to evaluate the efficacy of the guidelines. As the corpus scales, intra-annotator agreement will be computed on a re-annotated subset to validate the consistency and reliability of the framework.

6. Bibliographical References

R Grisham and B Sundheim. 1996. Message understanding: a brief history. In *Proceedings of the Sixth Message Understanding Conference*.

Milan Gritta. 2019. *Where are you talking about? advances and challenges of geographic analysis of text with application to disease monitoring*. University of Cambridge (United Kingdom).

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. 2017. Vancouver welcomes you! minimalist location metonymy resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1248–1259.

Mark Johnson and George Lakoff. 1980. *Metaphors we live by*, volume 1. University of Chicago press Chicago.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

Onkar Litake, Maithili Sabane, Parth Patil, Aparna Ranade, and Raviraj Joshi. 2023. Mono versus multilingual bert: A case study in hindi and marathi named entity recognition. In *Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2022*, pages 607–618. Springer.

Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. 2022. L3cube-mahaner: A marathi named entity recognition dataset and bert models. In *Proceedings of the WILDRE-6*

Workshop within the 13th Language Resources and Evaluation Conference, pages 29–34.

Katja Markert and Udo Hahn. 2002. Understanding metonymies in discourse. *Artificial intelligence*, 135(1-2):145–198.

Katja Markert and Malvina Nissim. 2002a. Metonymy resolution as a classification task. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 204–213.

Katja Markert and Malvina Nissim. 2002b. Towards a corpus annotated for metonymies: the case of location names. In *LREC*.

Katja Markert and Malvina Nissim. 2007. Semeval-2007 task 08: Metonymy resolution at semeval-2007. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 36–41.

Katja Markert and Malvina Nissim. 2009. Data and models for metonymy resolution. *Language Resources and Evaluation*, 43(2):123–138.

Marjorie McShane and Sergei Nirenburg. 2021. *Linguistics for the Age of AI*. Mit Press.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Malvina Nissim and Katja Markert. 2003. Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 56–63.

Nita Patil, Ajay Patil, and BV Pawar. 2020. Named entity recognition using conditional random fields. *Procedia Computer Science*, 167:1181–1188.

Nita Patil, Ajay S Patil, and BV Pawar. 2016. Issues and challenges in marathi named entity recognition. *International Journal on Natural Language Computing (IJNLC)*, 5(1):15–30.

Thierry Poibeau. 2006. Dealing with metonymic readings of named entities. *arXiv preprint cs/0607052*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.