# A Pipelined Approach to Entity-Aware Machine Translation

**Lola Kovalski and Ina Ocelli**

## Abstract

Entity-aware machine translation aims to enhance translation quality by explicitly addressing named entities, which often pose challenges for conventional translation systems. In this project, we propose a pipeline approach where named entities are identified and masked, translated separately to ensure contextual and semantic accuracy, and subsequently reintegrated into the translated text. Although our method focuses specifically on named entities, it fails to consistently outperform state-of-the-art translation systems. The results, while promising in certain cases, reveal limitations that highlight areas for improvement. We analyze these findings, examine factors contributing to the observed performance, and propose directions for future research to refine entity-aware translation.

## 1 Introduction

This paper introduces Entity-Aware Machine Translation (EA-MT), a system designed to leverage named entity recognition (NER) for improving the handling of proper nouns, places, and other culturally or domain-specific terms in translation. Machine translation systems frequently struggle with named entities, particularly rare or culturally unique ones, leading to errors that can alter the intended meaning of the text. To preserve semantic integrity and ensure accurate translation, named entities require special handling within the translation process.

For baseline comparison, we use the state-of-the-art Meta M2M100 model. We employ a pipelined approach where entities are identified using an NER model, masked, and translated separately. The masked sentence is then processed by Meta M2M100 model, after which the translated entities are reintegrated into the final output. By comparing our pipelined machine translation PMT) outputs with the Meta M2M100 (MT) outputs and the SemEval target translations (Target), we evaluate the effectiveness of our method.

To assess accuracy, we pursue a holistic approach, inspecting outputs by hand and calculating standard metrics (i.e. BLEU scores). This is motivated by Liang et al. (2024) who documented traditional metrics', like BLEU and COMET scores, failure to capture the nuance required to assess the accuracy and nuance of a translation involving entities.

Manual inspection reveals mixed results: while PMT and MT outputs are often more accurate than SemEval target translations, they are prone to distinct errors. PMT sometimes translates entities accurately but fails to reintegrate them correctly, introducing grammatical errors (e.g., incorrect articles). MT produces fewer grammatical issues but occasionally translates sentences in a literal word order, causing problems with certain syntactic structures in target languages (e.g., superlatives in Italian). SemEval target translations generally achieve the best results, particularly in translating idiomatic expressions, but PMT and MT frequently come close, with each method suffering from unique challenges. BLEU scores support these observations, indicating that while the MT model achieves higher n-gram overlap with reference translations, the performance gap between it and PMT is relatively narrow.

## 2 Related Work

Named entity translation has long been a challenging problem in machine translation due to the inherent complexity and low-frequency nature of named entities in training data. Prior research has proposed a variety of strategies to address these challenges, some of which align with the approach pursued in this work.

One significant line of research focuses on preprocessing techniques and entity masking to improve named entity translation. Mota et al. (2022)

developed a pipeline where named entities are identified and masked before translation, allowing the MT model to process the sentence without interference from unfamiliar entities. The masked entities are then translated separately and reintegrated into the final output. Their novel use of semantically equivalent placeholders in masking ensures compatibility with neural machine translation (NMT) models and reduces distortion in translation. Similarly, Sharma et al. (2023) employed a preprocessing step leveraging named entity recognition (NER) and transliteration to handle NE's before the translation process. Their system achieved high accuracy for translating person, location, and organization names, demonstrating the effectiveness of separating NE processing from general sentence translation.

Another important research direction involves enhancing NE translation through data augmentation and increased context diversity. Liang et al. (2024) observed that translation accuracy for NE's often depends on the frequency of the entity and the diversity of its surrounding context. To address this, they proposed a data augmentation strategy that generates synthetic sentences to improve context diversity and translation probability for low-frequency entities. Their experiments demonstrated significant improvements in general translation performance and NE-specific accuracy across multiple test sets. Tedeschi and Navigli (2022) contributed further to this effort with MultiNERD, a multilingual and fine-grained dataset designed to improve NER and NE disambiguation. By providing rich contextual information and supporting multiple languages, MultiNERD enables robust generalizations and accurate NE in NMT systems.

In addition to these preprocessing and data-focused strategies, hybrid approaches and the integrations of external knowledge sources have proven effective for NE translation. In adddition to preprocessing, Sharma et al. (2023) combined rules-based phoneme extraction with statistical and neural modeling techniques to improve transliteration, particularly for resource-poor languages. This hybrid approach highlights the potential of combining multiple methodologies to address the diverse challenges posed by NE's. Similarly, many studies have integrated external knowledge bases, such as Wikipedia and BabelNet, to enhance NE recognition and translation accuracy. These resources provide valuable information for training models on low-resource languages and domain-specific terminology.

Another innovative contribution to NE translation research is the development of entity-aware translation metrics. Liang et al. (2024) introduced a metric specific designed to evaluate the accuracy of NE translations. This metric complements traditional metrics like BLEU and better captures improvements in entity-specific translation performance.

Despite these advances, challenges remains. Issues such as gender agreement, transliteration errors, and context-dependent translations highlight the limitations of current approaches. For instance, while semantic masking strategies can improve NE handling, they may fail in highly ambiguous contexts, requiring fallback mechanisms that can degrade overall translation quality. Furthermore, maintaining linguistic agreement during reintegration of translated entities remains a persistent problem.

## 3 Dataset

Our original dataset was provided to us by the the SemEval-2025 Task organizers, but is a selective partition of a larger dataset called Mintaka (Sen et al., 2022). Mintaka is built off of Wikipedia and contains a list of questions and answers, with information on the entities present (such as span and type), and translation into several languages. Note that the Mintaka translations were generated by professional human translators. Furthermore, while the original intended use of Mintaka was for multilingual question and answer tasks, given the world and cultural knowledge implicit in the task of entity recognition and translation, its use here makes sense.

The SemEval partition contains a limited selection from this dataset and different set information per observation. Each line contains a source sentence, which is in English, a target sentence, which is in either Italian or Spanish in our case, and a list of entity codes present in the sentence. An entity code is of the form 'Q657' and in order to obtain the human language version of the entity, we search the Mintaka dataset and update the entity accordingly. In addition, we corrected labeling errors present in the SemEval dataset (i.e. language mislabeling, mixed source and target labels) to produce a clean, processed dataset.

We initially intended to use the Sem-Eval entity

labeling as our source of truth for the entity recognition and masking step, but preliminary manual inspection revealed potential issues. In particular, the SemEval data did not label all of the entities present in a sentences (Figure 1).

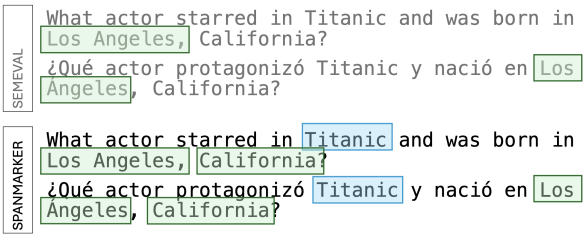


Figure 1: SemEval labeled entities (top) and NER labeled entities (bottom) for an English-Spanish translation instance

In Figure 1, the top translation highlights the entities labeled by SemEval and the bottom highlights the entities labeled by our NER model for the same translation instance. As you can see, SemEval only notes "Los Angeles" as a named entity whereas the NER model picks up on "Titantic," "Los Angeles," and "California." Given the recurrence of this phenomena throughout the dataset, we elected to use a state of the art NER model and its output as our source of entity truth when thinking about cascading error and masking.

## 4 Methodology

In contrast to the end-to-end approach to translation that most MT systems adopt, we divide the process into stages to handle named entities separately from the sentence in which they are situated. The hypothesis is that recognizing and translating entities independently reduces distortions in translation as the model might be less likely to bias translation (towards or away from the identity) or mistranslate. Our framework involves the following steps.

### 4.1 Named Entity Recognition (NER):

As discussed in the data section, rather than using the SemEval labeling, which links to Mintaka entity information, we use a pretrained NER model to identify and analyze identities. In particular, we use `tomararsen/span-marker-mbert-base-multinerd`, to identify named entities (e.g., PERSON, ORGANIZATION, LOCATION, etc) in the source text. We chose this model because it is trained on the MultiNERD dataset, the first language-agnostic methodology for auto-generating multilingual fine-grained annotations for NER and entity disambiguation. As of now, it is the top performing data production method for multilingual entity related tasks.

In terms of output, after running a sentence through this model, we receive a list of entities with several pieces of information. Figure 2 displays an example of output data that we hold onto for a single entity. See the Appendix ( A ) for the label set of potential types for this model.

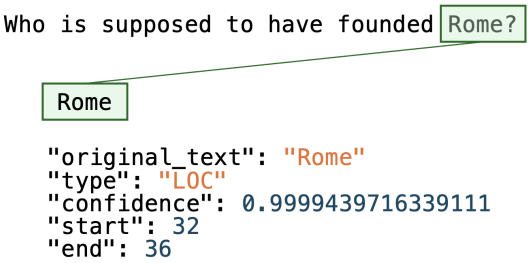


Figure 2: SpanMarker Output

### 4.2 Entity Masking:

At this stage, we replace identified named entities with placeholders that capture the pertinent information (e.g., `[ENTITY.PER.1]`), isolating them from the rest of the text.

If the NER model's confidence level is below 0.7 for a given entity – which was a minor for our data as less than 2% of observations fell in this range – the entity is not masked. If that is the only entity in a sentence, the translation should be identical to the end-to-end M2M100 translation.

### 4.3 Translation

Using the `M2M100ForConditionalGeneration` and `M2M100Tokenizer` models, we tokenize and then translate the masked sentence as seen in stage 2 of Figure 3. The model has demonstrated an ability to effectively tokenize our placeholders, but this might be an area for future exploration. Instead of using something that is clearly not a word, we might try using semantic equivalents (e.g. `[ENTITY.PER.1]` → "person") and compare the results.

With respect to the identified entity, we use a combination of logic and translation. We decide not to translate certain types, such as people, vehicles, instruments, plants, and media, instead integrating the untranslated entity into the translated sentence. See the Appendix ( A ) for more information on these types.

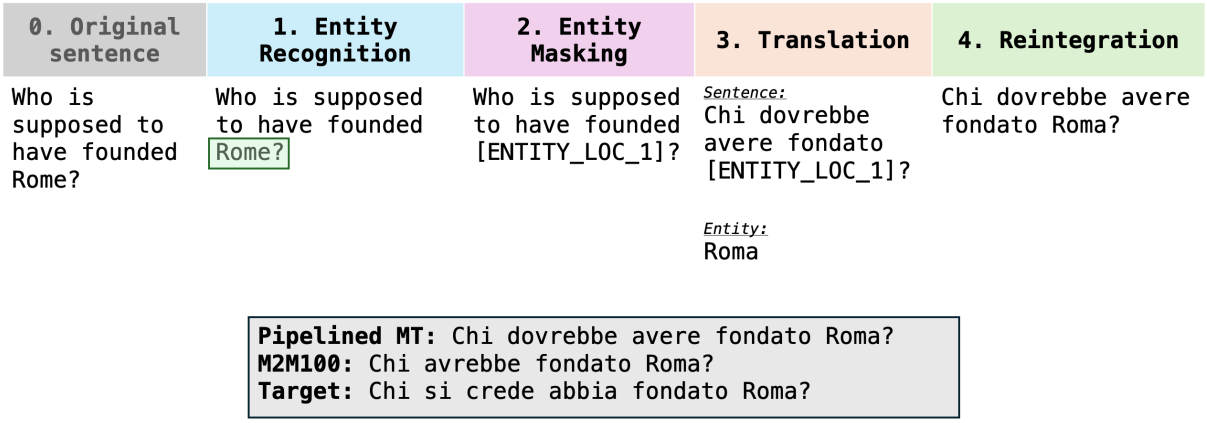


Figure 3: High level pipeline process on an English to Italian example

If an entity is not of these types, then we tokenize and translate with the M2M100 models as usual.

A potential area for improvement in this step would be to refine the translation logic. For instance, incorporating regex to identify additional entities, such as dates, phone numbers, and email addresses, could allow for more systematic handling during translation. Furthermore, the translation process requires additional contextual information. For example, while certain institutions might remain untranslated, there are cases where specific institutions must still be translated accurately. Addressing this challenge would involve gathering a more comprehensive and targeted training dataset. Another promising area for exploration is the translation of idioms, which can be regarded as unique entities. As idioms often lack literal translations, developing a translation model specifically trained on idiomatic expressions would be necessary.

### 4.4 Reintegration

After translation, placeholders in the text are replaced with the post-translation entity. This reintegration step uses a mapping of placeholders to the original entities, ensuring that each placeholder is replaced with the correct version (either translated or kept as-is depending on confidence of NER and type).

One of the challenges we run into with reintegration is ensuring sentence alignment with the entity. For instance, if the entity is gendered or plural, the pronouns in the masked sentences are not guaranteed to align. A way to tackle this problem might be to improve the information present in the mask.

### 4.5 Evaluation

To assess the accuracy of our translation, we use the both the target sentences from SemEval and translations from the M2M100 model.

The latter of which is state of the art for multilingual translation, and we use it to translate text from English to Italian and English to Spanish without any additional preprocessing or entity-specific handling. Together, the model highlights the strengths and limitations of separate entity processing, while the labeled target sentences allow us to evaluate the accuracy of machine translation more generally.

Manual inspection is crucial for evaluating translations because automated metrics like BLEU scores, often fail to capture the nuances of translation quality. For instance, a translation might achieve a high BLEU score despite being awkward or incorrect, simply by aligning with the reference text in superficial ways. Conversely, an accurate but idiomatic translation could receive a lower BLEU score if it uses different phrasing than the reference. To assess these subtleties, we discuss the contents of Table 1 and 2. The leftmost column is the SemEval, human labeled translation; the middle, our pipelined translation (PMT); the rightmost, M2M100 (M2M).

#### 4.5.1 English to Spanish Translations - Table 1

*Sentence 1: "Which war caused the most deaths, World War I or II?"*

Interestingly the target is the least literal translation, as it asks "Which war caused more deaths, the First of the Second World War?" The PMT and M2M model therefore more closely align to the original sentence. The difference between the two is translation of the entity, which shows that sep-

| Target | PipelinedMT | M2M100 |
|---|---|---|
| ¿Qué guerra causó más muertes, la Primera o la Segunda Guerra Mundial? | ¿Qué guerra causó la mayor cantidad de muertes, Guerra Mundial I o II? | ¿Qué guerra causó la mayor cantidad de muertes, la Primera o la Segunda Guerra Mundial? |
| ¿Dónde fue el primer ayuntamiento completamente femenino en los Estados Unidos? | ¿Dónde fue el primer consejo de la ciudad de todas las mujeres en la Estados Unidos? | ¿Dónde fue el primer consejo de la ciudad de todas las mujeres en los Estados Unidos? |
| ¿Cuántas veces se casó la actriz Bella de Crepúsculo? | ¿Cuántas veces se ha casado la actriz Bella de Twilight? | ¿Cuántas veces se ha casado la actriz Bella de Twilight? |
| ¿Qué país es mayor en superficie, Rusia o Canadá? | ¿Qué país es más grande en tamaño, Rusia o Canadá? | ¿Qué país es más grande en tamaño, Rusia o Canadá? |
| ¿En qué año se formó N Sync? | ¿Qué año se formó Nuevos? | ¿En qué año se formó NSYNC? |

Table 1: Selected Spanish translations for manual analysis

aration might have some effect on the translation of the entity itself. The pipelined model does not translate the numbers into words whereas the M2M model does. They maintain the same meaning, but since the source contains I and II, not "first" and "second," the pipelined translation is technically more correct.

*Sentence 2: "Where was the first all-woman city council in the United States?"*

This example again illustrates the increased accuracy of the MT models over the target. The target asks where the first all woman city hall is, which is distinct from the question as city hall is a building whereas city council is a group of people. What is different between the PMT and the M2M is the article preceding the entity. While the entity is translated the same, and correctly in both instances, the PMT has an incorrect article addressing it. We anticipated this being a problem because the translation of the masked sentence doesn't know the gender or plurarity of the entity that follows.

*Sentence 3: "How many times has the actress Bella from Twilight been married?"*

Again the MT outputs beat the target in terms of literal translation. The target asks the question in imperfect/simple past tense whereas the two MT sentences use perfect/past perfect tense (i.e. "was" versus "has"). In this example that distinction is not important, but, in others that might not be the case. Between the two MT models, however, they translate the entity "Twilight" differently. M2M directly translates whereas PMT keeps it as is. Both are valid as well known American movies, such as Twilight, are often referred to by their American title. The accuracy of this translation would likely depend on region and closeness to American culture.

*Sentence 4: What country is larger in size, Russia or Canada?*

All three models perform equally well on this example. The target translation uses a word that means roughly "area" whereas the MT sentences use a direct translation of "size," but this difference does not change meaning. All three translate the entities, Russia and Canada, correctly. Similarly, all three use the correct articles when addressing and have no problems with reintegration. This seems to be expected given the separation of the entity from the rest of the sentence via comma and the frequent appearance of country names (in contrast to niche cultural references) in text.

*Sentence 5: What year did NSYNC form?*

This example is one of the few examples where PMT fails to translate an identity. Despite the NER model identifying "NSYNC" as an organization in the source text with 94% confidence, its translation is inaccurate and fails to even convey the question.

A potential explanation for this is that NSYNC is not a formal word, but is also a band that was mildly popular twenty years ago. Therefore, it might not be reflected in enough training data for a model to recognize it on its own.

#### 4.5.2 English to Italian Translations - Table 2

*Sentence 1: "Which king of England was never officially crowned?"*

This translation is largely consistent across all three outputs. The target and MT translations

| Target | PipelinedMT | M2M100 |
|---|---|---|
| Quale re d'Inghilterra non è mai stato ufficialmente incoronato? | Quale re di Inghilterra non è mai stato coronato ufficialmente? | Quale re d'Inghilterra non è mai stato coronato ufficialmente? |
| Chi era il re di Gerusalemme che non poteva muovere le braccia per via della lebbra? | Chi era il re di Gerusalemme che non poteva spostare i suoi membri a causa della lepra? | Chi era il re di Gerusalemme che non poteva spostare i suoi membri a causa della lepra? |
| Chi è l'olimpico più decorato di tutti i tempi e nacque a Baltimora, MD? | Chi è il più decorato olimpico di tutti i tempi e nasce in Baltimore, MD?" | "Chi è il più decorato olimpico di tutti i tempi e nasce a Baltimore, MD? |
| Qual è stato il maggior terremoto mai registrato? | Qual è stato il primo terremoto mai registrato? | Qual è stato il primo terremoto mai registrato? |
| Il Monte Kilimangiaro si trova in Africa? | È Il Monte Kilimanjaro situato in Africa? | Il Monte Kilimanjaro si trova in Africa? |

Table 2: Selected Italian translations for manual analysis

are identical, while PMT introduces a minor error. Specifically, it uses the article "di" instead of "d' ", which occurs due to uncertainty during the first round translation. This uncertainty arises because the masked entity (in this case, England's translation, "Inghilterra") may begin with a vowel or consonant. This highlights the importance of providing the model with additional context for accurate grammatical handling of the entities. Alternatively, introducing a pipeline with additional layers or a "check" layer at the end could improve accuracy.

*Sentence 2: "Who was the king of Jerusalem who could not move his limbs due to leprosy?"*

The target translation is the closest to the intended meaning. Both PMT and MT translate "limbs" as "membri" (members of the body), which conveys the general idea but lacks precision. The target translation, on the other hand, uses the word "braccia" (arms), which is also not entirely accurate. The correct Italian term in this context should be "arti" (limbs). Additionally, the word "leprosy" is accurately translated as "lebbra" in the target, while PMT and MT opt for "lepra", a less common variant. This example suggests gaps in the training data for medical terms, anatomical specificity, and potentially other specialized fields in the MT models.

*Sentence 3: "Who is the most decorated Olympian of all time and was born in Baltimore, MD?"*

While all three translations may perform similarly on BLEU scores, the target translation correctly renders "l'olimpico più decorato", which follows the proper word order for superlatives in Italian. However, both PMT and MT follow the literal English word order, producing "il più decorato olimpico". This construction, though understandable, is not grammatically correct in Italian and reflects a lack of training on syntactic nuances. This example illustrates how literal translation (potentially following a literal word order) can lead to incorrect grammatical phrasing in Italian.

*Sentence 4: "What was the first largest earthquake ever recorded?"*

This sentence example lacks an entity, and we wanted to see how the models would perform on such a case. It results in identical outputs from PMT and MT. In terms of accuracy, the target translation omits "the first", focusing solely on "the largest earthquake". In contrast, both MT models emphasize "il primo" while neglecting "largest". This discrepancy may stem from Italian's tendency to treat "terremoto" (earthquake) as inherently significant, rendering "largest" redundant, which makes the MT models slightly more accurate in this case. It underscores the importance of aligning the model's focus with contextual nuances. *Sentence 5: "Is Mount Kilimanjaro located in Africa?"*

These translations are grammatically correct, but idiomatic expressions stand out. The target translation is the most idiomatic, using "si trova" (translation: where to find it?) which aligns with typical Italian phrasing for locations. PMT, while technically correct, uses "è situato" a translation that is

| Language | Model | BLEU-1 *(unigram)* | BLEU-2 *(bigram)* | BLEU-3 *(trigram)* | BLEU-4 *(4-gram)* |
|---|---|---|---|---|---|
| Italian | PipelineEAMT | 0.5686 | 0.4274 | 0.309 | 0.191 |
| | Meta M2M100 | 0.6248 | 0.5120 | 0.4048 | 0.3019 |
| Spanish | PipelineEAMT | 0.6598 | 0.5463 | 0.4426 | 0.3322 |
| | Meta M2M100 | 0.6972 | 0.6056 | 0.5218 | 0.4405 |

Table 3: BLEU Scores for Italian and Spanish

less common in everyday Italian when it comes to talking about locations. This example demonstrates how the lack of specific treatment for locations (or entities of different types) during translation with masking leads to overly literal outputs.

### 4.6 BLEU Scores - Table 3

We compute the BLEU scores for both our pipelined model and the Meta M2M100 model compared to labeled targets in Table 3. For both Italian and Spanish, we observe that the Meta M2M100 model consistently outperforms the PipelineEAMT across all n-gram levels.

For Italian, M2M100 achieves a BLEU-1 score of 0.6248 compared to 0.5686 for PipelineEAMT, and this rough magnitude of difference continues through BLEU-4. For Spanish, M2M100 outperforms again although the difference in less pronounced, except in the 4-gram case.

These results suggest that the M2M100 model aligns more closely with the reference translations. However, the gaps in performance are not drastic, especially for Spanish, where BLEU-1 difference is very small.

While BLEU scores are a widely used metric for evaluating translation quality, they come with significant limitations that should be considered. They measure the overlap of n-grams, which penalizes valid translations that differ in words and syntax, which can lead to low scores even when the translation is accurate. Relatedly, BLEU scores do not account for semantic equivalence or contextual appropriateness as a translation might have a high score but fail to preserve meaning in nuanced or idiomatic phrases.

Given these limitations and the negligible difference, especially in the Spanish case, the observed differences between PipelineEAMT and Meta M2M100 might not fully reflect their real-world translation quality.

## 5 Conclusions, Ethics, and Limitations

Entity-aware machine translation (EA-MT) represents a promising approach to addressing the unique challenges posed by named entities in translation tasks. Our pipeline approach, which integrates named entity recognition, masking, separate translation, and reintegration, demonstrates potential in improving translation accuracy for these critical components of text. Results from our analysis indicate that the method performs well in maintaining entity integrity in straightforward cases but struggles in more complex linguistic contexts. Reintegration errors and grammatical alignment issues are particularly evident when handling gendered or plural entities, and the translation of idiomatic expressions remains inconsistent.

Despite these promising results, the pipeline approach has limitations. Biases in training data, limited representation of rare entities, and the inherent difficulty of ensuring linguistic and cultural accuracy remain significant hurdles. Low-resource languages and dialects are particularly underrepresented, which impacts fairness and inclusivity. These challenges highlight the need for richer datasets, better handling of cultural nuances, and enhanced contextual understanding during translation. Additionally, our BLEU score analysis underscores the gap between our pipelined approach and state-of-the-art models like Meta M2M100, particularly in complex sentence structures and nuanced linguistic contexts.

While our pipelined method does not consistently outperform existing models, it highlights critical areas for further exploration, such as refining masking strategies, incorporating additional linguistic context, and developing methods to better address idiomatic expressions and grammatical agreements. Future work could also explore hybrid methods that integrate rule-based and neural approaches to address these persistent challenges.

Ultimately, advancing EA-MT requires technical

innovation and interdisciplinary collaboration to build robust datasets, refine evaluation metrics, and ensure fairness in language representation. This research lays a foundation for such efforts, offering insights into both the potential and the challenges in enhancing machine translation systems to handle named entities effectively.

## References


Tian Liang, Xing Wang, Mingming Yang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Addressing entity translation problem via translation difficulty and context diversity. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11628–11638.

Pedro Mota, Vera Cabarrao, and Eduardo Farah. 2022. Fast-paced improvements to named entity handling for neural machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 141–149, Ghent, Belgium. European Association for Machine Translation.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Radhika Sharma, Pragya Katyayan, and Nisheeth Joshi. 2023. Improving the quality of neural machine translation through proper translation of name entities. In *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, pages 1–4. IEEE.

Simone Tedeschi and Roberto Navigli. 2022. Multinerd: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812.


# A Appendix

| Class | Description | Examples |
|---|---|---|
| PER (person) | People | Ray Charles, Jessica Alba, Leonardo DiCaprio, Roger Federer, Anna Massey. |
| ORG (organization) | Associations, companies, agencies, institutions, nationalities and religious or political groups | University of Edinburgh, San Francisco Giants, Google, Democratic Party. |
| LOC (location) | Physical locations (e.g. mountains, bodies of water), geopolitical entities (e.g. cities, states), and facilities (e.g. bridges, buildings, airports). | Rome, Lake Paiku, Chrysler Building, Mount Rushmore, Mississippi River. |
| ANIM (animal) | Breeds of dogs, cats and other animals, including their scientific names. | Maine Coon, African Wild Dog, Great White Shark, New Zealand Bellbird. |
| BIO (biological) | Genus of fungus, bacteria and protoctists, families of viruses, and other biological entities. | Herpes Simplex Virus, Escherichia Coli, Salmonella, Bacillus Anthracis. |
| CEL (celestial) | Planets, stars, asteroids, comets, nebulae, galaxies and other astronomical objects. | Sun, Neptune, Asteroid 187 Lamberta, Proxima Centauri, V838 Monocerotis. |
| DIS (disease) | Physical, mental, infectious, non-infectious, deficiency, inherited, degenerative, social and self-inflicted diseases. | Alzheimer's Disease, Cystic Fibrosis, Dilated Cardiomyopathy, Arthritis. |
| EVE (event) | Sport events, battles, wars and other events. | American Civil War, 2003 Wimbledon Championships, Cannes Film Festival. |
| FOOD (food) | Foods and drinks. | Carbonara, Sangiovese, Cheddar Beer Fondue, Pizza Margherita. |
| INST (instrument) | Technological instruments, mechanical instruments, musical instruments, and other tools. | Spitzer Space Telescope, Commodore 64, Skype, Apple Watch, Fender Stratocaster. |
| MEDIA (media) | Titles of films, books, magazines, songs and albums, fictional characters and languages. | Forbes, American Psycho, Kiss Me Once, Twin Peaks, Disney Adventures. |
| PLANT (plant) | Types of trees, flowers, and other plants, including their scientific names. | Salix, Quercus Petraea, Douglas Fir, Forsythia, Artemisia Maritima. |
| MYTH (mythological) | Mythological and religious entities. | Apollo, Persephone, Aphrodite, Saint Peter, Pope Gregory I, Hercules. |
| TIME (time) | Specific and well-defined time intervals, such as eras, historical periods, centuries, years and important days. No months and days of the week. | Renaissance, Middle Ages, Christmas, Great Depression, 17th Century, 2012. |
| VEHI (vehicle) | Cars, motorcycles and other vehicles. | Ferrari Testarossa, Suzuki Jimny, Honda CR-X, Boeing 747, Fairey Fulmar. |

Figure 4: Label Set for `tomararsen/span-marker-mbert-base-multinerd`