Benchmarking Bilingual Text Anonymization and Automatic Term Extraction Approaches

Angela Forzatti

Abstract

This study presents a comparative evaluation of three Natural Language Processing (NLP) approaches — Conditional Random Fields (CRF), Bidirectional Long Short-Term Memory networks with a CRF layer (BiLSTM-CRF), and fine-tuned Bidirectional Encoder Representations from Transformers (BERT) — applied to two sequence labelling tasks: Text Anonymization via Named Entity Recognition (NER) and Automatic Term Extraction (ATE). The models were trained and tested on multiple languages, specifically English and Spanish for NER, and English and Italian for ATE, utilizing established datasets such as the CoNLL 2002/2003 and the BitterCorpus. Experimental results indicate that the fine-tuned BERT model achieves superior performance in NER-based Text Anonymization, whereas the CRF model outperforms deep learning architectures on ATE tasks. These findings highlight the varying suitability of traditional machine learning and transformer-based methods depending on task type, language, and dataset characteristics. The presented benchmarks provide a foundational framework for developing more complex systems, which can be further optimized and adapted to specific requirements.

1 Introduction and background

This study compares three Natural Language Processing (NLP) approaches for sequence labelling tasks in a bilingual setting: classical machine learning (Conditional Random Fields, CRF), deep learning architectures (Bidirectional Long Short-Term Memory networks with CRF layer, BiLSTM-CRF), and state-of-the-art transformer models (fine-tuned BERT). The aim is to clarify the relative strengths and limitations of each methodology under different conditions. The models, selected in line with prior literature [Tran et al., 2023] [Asimopoulos et al., 2024], were applied consistently across both tasks to emphasize comparability; however, the datasets differ markedly in size and quality, reflecting practical constraints that limited the identification of perfectly comparable corpora.

The first task, Text Anonymization, uses Named Entity Recognition (NER) to detect and mask Personally Identifiable Information (PII) for privacy and regulatory compliance [European Parliament and Council of the European Union, 2016]. Since NER models classify entities into categories (such as person, location, organization, etc.), the categorization approach of substituting an entity with a categorical label was applied [Asimopoulos et al., 2024]. The main challenge of this task lies in accurately identifying these entities to effectively obfuscate sensitive data while preserving the overall utility of the text.

The second task, Automatic Term Extraction (ATE), involves the automatic identification of domain-specific terminology from textual corpora for applications such as technical translation or terminology management [Tran et al., 2023]. This task struggles with the inherent difficulty of defining what constitutes a ”term”; they are generally described as ”textual expressions that denote concepts within a specific field of expertise” [Tran et al., 2023], but their boundaries and characteristics often vary by domain and context.

Recent systematic surveys [Tran et al., 2023] [Asimopoulos et al., 2024] have traced the evolution of several approaches to NER and ATE tasks during the years. Early approaches relied on rule-based systems and handcrafted features [Tran et al., 2023]. The introduction of machine learning models, such as CRF, improved precision and generalizability thanks to feature engineering techniques that allowed models to learn associations between features and labels, but required substantial manual effort [Tran et al., 2023] [Asimopoulos et al., 2024] [Judea et al., 2014]. More recent developments in deep learning, such as BiLSTM-CRF architectures, have introduced the capacity to capture bidirectional contextual dependencies and sequential information without extensive feature engineering [Tran et al., 2023]; [Asimopoulos et al., 2024] [Huang et al., 2015]. Transformer-based models, such as BERT, have further advance performance by leveraging rich contextual embeddings that capture syntactic and semantic information at multiple levels [Devlin et al., 2019]; fine-tuning BERT on task-specific data has yielded state-of-the-art results across a variety of NLP tasks [Tran et al., 2023] [Asimopoulos et al., 2024].

2 Methodology

2.1 Datasets

For the Named Entity Recognition task, the CoNLL 2003 dataset [Sang and Meulder, 2003] was initially considered due to its widespread use and high-quality annotations for English and German. However, due to the unavailability of the German portion, the CoNLL 2002 Spanish dataset [Tjong Kim Sang, 2002] was selected to complement the English data while maintaining consistency within the CoNLL dataset family. Both datasets were sourced from HuggingFace repositories [1] [2] and provide annotated entities based on the BIO tagging scheme across four categories: PERSON, LOCATION, ORGANIZATION, and MISC. Additionally, part-of-speech (POS) tags are available for each language.

For the Automatic Term Extraction task, the BitterCorpus dataset [Arčan et al., 2014] was used: the GNOME subset comprises 55 parallel documents pertaining to the Information Technology (IT) domain with 313 Italian and 282 English annotated terms. BitterCorpus is not directly accessible via repositories and required manual download[3] and upload. Furthermore, the term annotations are provided in separate XML files, necessitating a custom preprocessing step to align the annotations with the corresponding text segments and assign the correct BIO labels.

Corpus	Total tokens
CoNLL 2002 Spanish	369171
CoNLL 2003 English	301418
BitterCorpus English	4214
BitterCorpus Italian	4331

Table 1: Total tokens per corpus.

2.2 Label conventions

Both Named Entity Recognition and Automatic Term Extraction tasks employed the widely adopted BIO tagging scheme for sequence labelling [Ramshaw and Marcus, 1995]. The BIO scheme is a standard approach for handling multi-word expressions in token-level annotation, where each token is classified as Beginning of an entity (B), Inside an entity (I), or Outside of any entity (O). In the context of the NER task, entity types were annotated with BIO tags corresponding to the four main categories identified by the CoNLL datasets: B-PER or I-PER for people, B-LOC or I-LOC for locations, B-ORG or I-ORG for organizations, B-MISC or I-MISC for miscellaneous entities, and O for tokens outside any named entity. For the ATE task, the BIO scheme was simplified to distinguish terms from non-terms. Tokens belonging to a domain-specific term were annotated as B-TERM if they were the first token in a term, I-TERM if they were inside a multi-token term, and O otherwise.

2.3 Models

Three models were trained and compared for both tasks:

- **Conditional Random Fields (CRF)**: a probabilistic model well-suited for sequence labelling tasks introduced by [Lafferty et al., 2001] . The model ”encodes a conditional probability distribution with a given set of features” [Zhang et al., 2008] and effectively captures neighbouring tag information at the sentence level [Huang et al., 2015].

- **BiLSTM-CRF:** a hybrid model introduced by [Huang et al., 2015] which combines a bidirectional Long Short-Term Memory (BiLSTM) network that captures both ”past and future input features” and a CRF output layer that leverages ”sentence level tag information”, ensuring a logical progression between tags.

[1] https://huggingface.co/datasets/eriktks/conll2002

[2] https://huggingface.co/datasets/eriktks/conll2003

[3] https://mt.fbk.eu/bittercorpus/

- **Fine-tuned BERT:** BERT, introduced by [Devlin et al., 2019], is a deep contextualized language model based on the Transformer architecture [Vaswani et al., 2023]. For this study, BERT was fine-tuned for token classification, leveraging its abilities to learn bidirectional context representations without hand-crafted features [Hazem et al., 2022].

3 Implementation[4][5]

3.1 Named Entity Recognition

The **CRF** model was adapted from the official sklearn-crfsuite tutorial [Korobov, 2015]. SpaCy [Honnibal et al., 2020] was used to standardize POS tagging across the two languages, ensuring cross-lingual consistency and avoiding feature mismatch issues during training. Linguistic, syntactic and contextual features were extracted for each token and its immediate neighbours, along with sentence boundaries. These features were then fed to the CRF model with "L-BFGS training algorithm (it is default) with Elastic Net (L1 + L2) regularization" [Korobov, 2015] for training and evaluation.

The **BiLSTM-CRF** architecture, adapted from [Huang et al., 2015], followed a similar preprocessing pipeline to the CRF, though the handcrafted features were heavily reduced to a few linguistic, morphological and contextual ones, to avoid overwhelming the model. The BiLSTM layer embeds words and features and decodes the most likely label sequences using a CRF layer. The model is trained over multiple epochs using the Adam optimizer and evaluated.

The **BERT** `bert-base-multilingual-cased` [Devlin et al., 2019] was accessed through HuggingFace[6] and fine-tuned following the official tutorial for token classification[7]. Training was handled with the Trainer API for 3 epochs, with the checkpoint achieving the highest validation F1 saved for evaluation on the test set.

For each model, inference was run by loading the trained model, preprocessing the sample text similarly to how the training dataset was preprocessed, then replacing each predicted entity with their corresponding labels.

3.2 Automatic Term Extraction

The **CRF** model was inspired by the one developed by [Zhang et al., 2008] for keyword extraction, excluding positional features (which are more suited for keyword identification rather than term). The BitterCorpus was loaded and each document was annotated using the gold-standard terms from the XML and the BIO tagging scheme. Features, including lexical, morphological, syntactic, and contextual cues, as well as TF-IDF and term frequency values, were extracted per token. The model was trained using the `lbfgs` algorithm and evaluated.

The **BiLSTM-CRF** architecture followed the one used for the NER model, adjusting the dataset loading and processing to the BitterCorpus structure. The feature engineering portion was also heavily reduced from the CRF setup to only a few basic lexical, morphological and contextual features. The model is trained for 80 epochs and evaluated.

The **BERT** `bert-base-multilingual-cased` was trained using the same pipeline from the official tutorial. As per the other models, the BitterCorpus needed a custom preprocessing pipeline. The model was trained for 10 epochs, with the checkpoint achieving the highest validation F1 (epoch 7), saved for evaluation.

Similarly to the inference run for the NER models, the saved models were loaded, the sample texts preprocessed and then a function would detect the terms and extract them as a simple list.

4 Evaluation and results

4.1 On test dataset

Model performance was evaluated on the test split of each dataset using precision, recall, and F1-score metrics (in the following tables, only the weighted average for each metric is reported). The metrics explicitly exclude "O" tags, as its dominance would have heavily skewed the results. Detailed per-tag metrics are provided in the Appendix.

[4] A folder is provided containing the models' code, inference results and the trained models themselves.

[5] The code was designed and written with the help of AI systems (Claude and ChatGPT).

[6] https://huggingface.co/google-bert/bert-base-multilingual-cased

[7] https://huggingface.co/docs/transformers/tasks/token_classification

4.1.1 NER and anonymization

As shown in Table 2, the fine-tuned multilingual BERT model substantially outperformed both CRF and BiLSTM-CRF models across all metrics. It achieved an F1-score of 0.90, with high precision (0.89) and recall (0.90), demonstrating strong capabilities in detecting sensitive entities. In contrast, both the CRF and BiLSTM-CRF models yielded moderate results, though the CRF model performed slightly better than the BiLSTM-CRF model.

Model	Precision	Recall	F1
CRF	0.79	0.77	0.78
BiLSTM-CRF	0.78	0.74	0.76
BERT	**0.89**	**0.90**	**0.90**

Table 2: Performance metrics for NER/anonymization.

4.1.2 ATE

Table 3 summarizes the evaluation metrics for the three ATE models. The CRF model achieved the highest precision (0.77) and F1 score (0.69), though not by much; while BERT achieved the best recall (0.71) but much lower precision. The BiLSTM-CRF performed moderately as it concerns precision and recall.

Model	Precision	Recall	F1
CRF	**0.77**	0.63	**0.69**
BiLSTM-CRF	0.68	0.65	0.64
BERT	0.61	**0.71**	0.66

Table 3: Performance metrics for ATE

4.2 On simulated real-world texts

Inference was conducted on simulated real-world texts[8], designed to reflect practical anonymization and term extraction scenarios, the latter in both in-domain and out-of-domain contexts. This approach aimed at measuring each model's generalizability and effectiveness in applied contexts.

The following tables display the numbers of True Positives (TP), False Positives (FP) and False Negatives (FN) for each model. Entities that were partially identified (i.e. correct boundaries but wrong labels or wrong boundaries but correct labels or wrong boundaries and wrong labels) were evaluated using a strict approach, where they each contribute to both False Positives and False Negatives [Esuli and Sebastiani, 2010] [9]. Finally, a manual calculation of precision, recall and F1 score is provided to assess if the models' performance in real-world applications is similar to the one observed on the test set. The full texts, with the manually annotated entries, are provided in the Appendix, while the model's annotations can be found in the notebooks.

4.2.1 NER and anonymization

For the English text, BERT achieved perfect recall (15 out of 15 terms) and the highest F1 (0.90), misclassifying only a few acronyms as organizations; a mistake made by the other two models as well. Mirroring the test performance, the CRF model slightly outperformed the BiLSTM-CRF, which tended to over-identify entities, especially ORG. For Spanish, performance decreased across all models, particularly for BERT, whose recall dropped to 0.78 (missing 4 out of 19 entities). Nonetheless, it still performed slightly better than the CRF and BiLSTM-CRF.

Model	TP	FN	FP	Precision	Recall	F1
CRF	12	3	5	0.70	0.80	0.74
BiLSTM-CRF	12	3	7	0.63	0.80	0.70
BERT	15	0	3	**0.83**	**1**	**0.90**

Table 4: Inference on English text

[8] Simulated by ChatGPT.

[9] Though out of the scope of this study, a more lenient approach towards partial matches, where the evaluation is carried out per-token (and not per-entity) [Esuli and Sebastiani, 2010], could be applied for comparison.

Model	TP	FN	FP	Precision	Recall	F1
CRF	14	5	6	0.70	0.73	0.71
BiLSTM-CRF	14	5	7	0.66	0.73	0.69
BERT	15	4	5	**0.75**	**0.78**	**0.76**

Table 5: Inference on Spanish text

4.2.2 ATE

The in-domain texts are simulated paragraphs on Information Technology themes. For English, the results followed the same trend as the test data: CRF achieved the highest precision, identifying only correct terms, whereas BERT demonstrated higher recall (10 out of 12 terms). For Italian, BERT outperformed all models in all metrics, achieving perfect recall (6 out of 6 terms), though precision was quite low (only 0.54). The BiLSTM-CRF, on the other hand, consistently underperformed, particularly in English, where it achieved an F1 score of 0.27; the model struggled with False Positives and word boundaries, often grouping together long strings of words.

Model	TP	FN	FP	Precision	Recall	F1
CRF	9	3	0	**1.00**	0.75	**0.85**
BiLSTM-CRF	3	9	7	0.30	0.25	0.27
BERT	10	2	2	0.83	**0.83**	0.83

Table 6: Inference on English in-domain text

Model	TP	FN	FP	Precision	Recall	F1
CRF	3	3	5	0.37	0.50	0.42
BiLSTM-CRF	3	3	9	0.25	0.50	0.33
BERT	6	0	5	**0.54**	**1.00**	**0.70**

Table 7: Inference on Italian in-domain text.

The outside-of-domain texts pertain to the Economics domain for English and to the Medical domain for Italian. In this case, the models did not perform satisfactorily; however, they all performed quite similarly to each other, deviating from the trend observed for the in-domain texts. For English, the CRF achieves the overall highest performance, thanks to the highest precision (only 3 False Positives), while in Italian the BiLSTM-CRF has the highest recall (3 terms out of 11), but identifies too many False Positives, bringing down its precision dramatically.

Model	TP	FN	FP	Precision	Recall	F1
CRF	2	7	3	**0.40**	**0.22**	**0.28**
BiLSTM-CRF	2	7	4	0.33	**0.22**	0.26
BERT	2	7	8	0.20	**0.22**	0.21

Table 8: Inference on English out-of-domain text.

Model	TP	FN	FP	Precision	Recall	F1
CRF	2	9	1	**0.66**	0.18	**0.28**
BiLSTM-CRF	3	8	11	0.21	**0.27**	0.24
BERT	2	9	1	**0.66**	0.18	**0.28**

Table 9: Inference on Italian out-of-domain text

5 Discussion and future works

The experimental results highlight the distinct strengths and limitations across the evaluated approaches, emphasizing the importance of selecting appropriate methods based on task characteristics, language, dataset availability and other contextual requirements.

Overall, all three models performed better in the Named Entity Recognition task compared to the Automatic Term Extraction task. This disparity may primarily be due to two main factors: (1) quality and size of the datasets and (2) the variability and complexity of domain terminology, as mentioned in the introduction. Regarding the first factor, the quality and size of the CoNLL 2002/2003 datasets appear to have played a crucial role in effectively fine-tuning BERT, a model considered state-of-the-art when it comes to a variety of NLP tasks. Notably, one of its known limitations is its training instability and sensitivity to hyperparameters with small datasets [Devlin et al., 2019], such as the one used for ATE. To validate this hypothesis, future experiments should perform a comparative analysis either by applying oversampling techniques to the ATE dataset or employing a larger dataset altogether. In contrast, the CRF model proved more suitable for smaller datasets, as the richness of the handcrafted features helped compensate for the limited availability of training data. The BiLSTM-CRF model, however, underperformed in both tasks, particularly in ATE, pointing to the conclusion that combining a smaller dataset with a reduced feature set is not ideal.

In practical applications, the NER models have achieved quite satisfactory results. However, a key limitation remains: the entities annotated in standard NER datasets do not necessarily correspond to Personally Identifiable Information (PII), which is the type of information that the [European Parliament and Council of the European Union, 2016] deem should be protected and obfuscated. Some datasets specifically dealing with PII[10] could be tested to improve Text Anonymization. Alternatively, custom datasets could be developed to capture the PII types of interest, according to specific needs and requirements; however, this process is resource- and time-intensive. On the other hand, the difference between the ATE models' performance on in-domain and out-of-domain text is remarkable and highlights the struggle to generalize across domains, supporting the hypothesis that the variability of domain terminology plays a big role. This inference step further highlighted performance differences across languages, with English consistently outperforming both Spanish and Italian (at least in-domain). This result is quite unexpected, as both CRF and BiLSTM-CRF are language-agnostic models and BERT was specifically chosen in its multilingual version to better support bilingual fine-tuning. A possible explanation could lie in the intrinsic linguistic characteristics of Romance languages compared to English or in the way model architectures have coincidentally been optimized for English (e.g. handcrafted features). Further dedicated experiments and a more in-depth literature review would be needed to verify these hypotheses, which are beyond the scope of this study.

While some exploratory experiments were performed during development to refine the training setup, the final results reported here are based on a single end-to-end pipeline for each approach. As a consequence, statistical testing on variability cannot be conducted on this study. Nevertheless, such testing would be valuable to determine the significance of the observed models' performance across tasks and evaluation conditions.

Future research could focus on different issues. First, the development or integration of dedicated PII datasets could significantly enhance the applicability of NER models for Text Anonymization, while domain adaptation techniques may improve the generalizability of ATE models. Second, exploring data augmentation and oversampling methods could help address the scarcity of high-quality annotated datasets, particularly for under-resourced languages, domains or tasks. Finally, leveraging Large Language Models (LLMs) in zero-shot or few-shot settings, may offer a resource-efficient alternative for both Automatic Term Extraction and Anonymization tasks.

References

[Arčan et al., 2014] Arčan, M., Turchi, M., Tonelli, S., and Buitelaar, P. (2014). Enhancing statistical machine translation with bilingual terminology in a cat environment.

[Asimopoulos et al., 2024] Asimopoulos, D., Siniosoglou, I., Argyriou, V., Karamitsou, T., Fountoukidis, E., Goudos, S. K., Moscholios, I. D., Psannis, K. E., and Sarigiannidis, P. (2024). Benchmarking advanced text anonymisation methods: A comparative study on novel and traditional approaches.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Esuli and Sebastiani, 2010] Esuli, A. and Sebastiani, F. (2010). Evaluating information extraction. In *Multilingual and Multimodal Information Access Evaluation*, pages 100–111. Springer Berlin Heidelberg.

[10] For example: https://huggingface.co/datasets/gretelai/synthetic_pii_finance_multilingual

[European Parliament and Council of the European Union, 2016] European Parliament and Council of the European Union (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation).

[Hazem et al., 2022] Hazem, A., Bouhandi, M., Boudin, F., and Daille, B. (2022). Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 648–662. European Language Resources Association.

[Honnibal et al., 2020] Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spacy: Industrial-strength natural language processing in python.

[Huang et al., 2015] Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging.

[Judea et al., 2014] Judea, A., Schütze, H., and Bruegmann, S. (2014). Unsupervised training set generation for automatic acquisition of technical terminology in patents. In Tsujii, J. and Hajic, J., editors, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 290–300. Dublin City University and Association for Computational Linguistics.

[Korobov, 2015] Korobov, M. (2015). Let's use conll 2002 data to build a ner system.

[Lafferty et al., 2001] Lafferty, J., Mccallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

[Ramshaw and Marcus, 1995] Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

[Sang and Meulder, 2003] Sang, E. F. T. K. and Meulder, F. D. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition.

[Tjong Kim Sang, 2002] Tjong Kim Sang, E. F. (2002). Introduction to the conll-2002 shared task: Language-independent named entity recognition.

[Tran et al., 2023] Tran, H. T. H., Martinc, M., Caporusso, J., Doucet, A., and Pollak, S. (2023). The recent advances in automatic term extraction: A survey.

[Vaswani et al., 2023] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

[Zhang et al., 2008] Zhang, C., Wang, H., Liu, Y., Wu, D., Liao, Y., and Wang, B. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, 4:1169–1180.

6 Appendix

6.1 Detailed per-entity evaluation metrics

The following tables summarize in more detail the evaluation metrics per entity for the NER models and the evaluation metrics per tag for the ATE models (excluding for BERT, as the evaluation averages B-TERM and I-TERM in TERM):

Entity	Model	Precision	Recall	F1
LOC	CRF	0.74	0.68	0.71
	BiLSTM-CRF	0.82	0.60	0.69
	BERT	**0.91**	**0.91**	**0.91**
MISC	CRF	0.69	0.63	0.66
	BiLSTM-CRF	0.66	0.61	0.63
	BERT	**0.74**	**0.79**	**0.76**
ORG	CRF	0.77	0.76	0.77
	BiLSTM-CRF	0.73	0.76	0.74
	BERT	**0.87**	**0.90**	**0.88**
PER	CRF	0.85	0.88	0.87
	BiLSTM-CRF	0.85	0.87	0.86
	BERT	**0.96**	**0.96**	**0.96**

Table 10: Performance metrics for different models by entity type

Tag	Model	Precision	Recall	F1
B-TERM	CRF	**0.72**	**0.61**	**0.66**
	BiLSTM-CRF	0.66	0.59	0.62
I-TERM	CRF	**0.81**	**0.64**	**0.72**
	BiLSTM-CRF	0.73	0.59	0.65

Table 11: Performance metrics for different models by tag

6.2 Evaluation texts

The following texts were used to run inference on the NER and ATE models. The entities have been annotated manually to compare the performance of the models. The models' annotations can be found in the respective notebooks.

English text for NER:

> On October 12, 2023, the (Global Tech Summit)[MISC] kicked off in (San Francisco, California)[LOC], drawing over 5,000 attendees from around the world. The event was hosted by (TechNova Inc.)[ORG], a leading software company headquartered in (Berlin, Germany)[LOC]. Keynote speaker Dr. (Aisha Rahman)[PER], CTO of (QuantumLeap Technologies)[ORG], unveiled their latest product: the QL-9000, a quantum processor capable of performing 10 trillion operations per second. She emphasized the importance of ethical AI development and announced a partnership with the (United Nations)[ORG] to promote responsible tech practices. Other notable speakers included (Elon Zhang)[PER], founder of (GreenGrid Energy)[ORG], and (Maria Gonzalez)[PER], a data scientist at the (University of São Paulo)[ORG]. Panels covered topics ranging from cybersecurity in the age of AI to the future of wearable health devices. The summit concluded with a gala dinner at the (Ritz-Carlton)[LOC], where attendees networked with executives from companies like (Microsoft)[ORG], (Alibaba)[ORG], and (Tata Consultancy Services)[ORG].

Spanish text for NER:

> El 5 de septiembre de 2025, líderes de la (Unión Europea)[ORG] se reunieron en (Bruselas)[LOC] para discutir nuevas estrategias energéticas. La presidenta de la (Comisión Europea)[ORG], (Ursula von der Leyen)[PER], presentó un plan de inversión de 300 mil millones de euros destinado a acelerar la transición hacia energías renovables. Durante la conferencia, representantes de países como (Alemania)[LOC], (Francia)[LOC], (Italia)[LOC] y (España)[LOC] expresaron su apoyo al proyecto, que será coordinado por la (Agencia Europea de Energía)[ORG]. El ministro de Medio Ambiente de (Alemania)[LOC], (Klaus Richter)[PER], destacó la importancia de reducir la dependencia del gas natural ruso. También estuvieron presentes delegados de empresas como (Siemens)[ORG], (Iberdrola)[ORG] y (TotalEnergies)[ORG], quienes ofrecieron propuestas tecnológicas para mejorar la eficiencia energética en zonas rurales. El evento fue cubierto por medios como (El País)[ORG], (Le Monde)[ORG] y (Deutsche Welle)[ORG]. La próxima reunión está programada para el 20 de noviembre

en (Viena)[LOC], donde se espera la participación de expertos del (Instituto Internacional de Energía Verde)[ORG].

In-domain English text for ATE:

(Blockchain technology)[TERM] provides a (decentralized ledger)[TERM] that enhances (transaction security)[TERM] and (transparency)[TERM]. (Smart contracts)[TERM] allow automated agreements without the need for intermediaries, reducing costs and delays. (Distributed ledger systems)[TERM] improve traceability across (supply chains)[TERM]. Furthermore, (encryption protocols)[TERM] safeguard (user data)[TERM] against unauthorized access and (cyber threats)[TERM], ensuring (data privacy)[TERM] in increasingly connected (networks)[TERM].

In-domain Italian text for ATE:

I (sistemi di intelligenza artificiale)[TERM] stanno trasformando profondamente il settore manifatturiero attraverso l'(automazione avanzata)[TERM]. Gli (algoritmi di apprendimento automatico)[TERM] analizzano grandi volumi di dati per ottimizzare i processi produttivi. Le (reti neurali artificiali)[TERM] simulano il funzionamento del cervello umano per migliorare il riconoscimento delle immagini e del linguaggio. L'elaborazione del (linguaggio naturale)[TERM] consente un'interazione più efficace tra uomo e macchina, aprendo nuove possibilità nell'assistenza clienti e nella (robotica)[TERM].

Outside of domain English text for ATE:

The (central bank)[TERM]'s decision to raise (interest rates)[TERM] was aimed at curbing rising (inflation)[TERM], which has impacted (consumer purchasing power.)[TERM] Investors are increasingly turning to (portfolio diversification strategies)[TERM] to mitigate risk amid volatile markets. (Stock market fluctuations)[TERM] have been exacerbated by geopolitical tensions and changes in (trade policies)[TERM]. Financial analysts continue to closely monitor quarterly (earnings reports)[TERM] and (macroeconomic indicators)[TERM] to advise their clients effectively.

Outside of domain Italian text for ATE:

L'ospedale ha implementato un nuovo protocollo per la gestione dei pazienti con (insufficienza cardiaca cronica)[TERM], mirato a ridurre i (ricoveri ospedalieri)[TERM]. Il (trattamento farmacologico)[TERM] prevede l'uso combinato di (beta-bloccanti)[TERM], (ACE-inibitori)[TERM] e (diuretici)[TERM] per migliorare la (funzione cardiaca)[TERM] e alleviare i sintomi. La (riabilitazione cardiaca)[TERM], inclusa l'(attività fisica supervisionata)[TERM] e la (consulenza nutrizionale)[TERM], è fondamentale per migliorare la qualità di vita dei pazienti. Inoltre, il monitoraggio continuo tramite (dispositivi indossabili)[TERM] permette di rilevare precocemente segni di peggioramento e intervenire tempestivamente.