Document Level Information Extraction in Low-Resource Languages

Mikel Zubillaga

University of the Basque Country UPV/EHU, HiTZ Basque Center for Language Technology - Ixa NLP Group

Abstract

This thesis addresses two key limitations in current Information Extraction (IE) systems: their inability to process text beyond individual sentences and their poor performance on low-resource languages like Basque. We propose using Large Language Models to enable document level information extraction while developing knowledge transfer techniques to improve results for languages with limited data. By combining these approaches, we aim to significantly enhance the quality of IE in documents, particularly for complex content and underrepresented languages, contributing to the advancement of document-level and multilingual information extraction.

Keywords

Information Extraction (IE), Large Language Models (LLMs), Document level NLP, Cross-lingual Transfer Learning, Low-resource Languages

1. Reason for the proposed research

In today's digital era, the majority of data exists as raw text, making Information Extraction (IE) essential for managing textual data floods and building effective applications. While deep learning techniques have established the state of the art across most language technology tasks, including IE, they face significant limitations. These systems struggle with long-form content since they operate only at the sentence level, creating coherence problems and reducing the quality of extracted information when sentences are processed in isolation. Furthermore, their dependence on large training datasets results in poor performance for low-resource languages like Basque. This thesis addresses these challenges through a dual approach: first, utilizing Large Language Models (LLMs) to extend IE capabilities beyond sentences to entire documents or document collections; second, investigating knowledge transfer techniques to improve performance for languages with limited data resources. We hypothesize that combining these approaches will significantly enhance the quality of extracted information from documents, advancing the current state of the art in information extraction technology.

2. Background and related work

In recent years, the amount of text created by humans has increased exponentially. These texts —news, social media messages, blogs, etc.— contain a lot of valuable information, but their processing has become impossible for human experts alone. Techniques capable of automatically extracting valuable information from textual corpora of this scale are researched in the field of Information Extraction (IE). These techniques can be found behind current cutting-edge technologies, such as Google Knowledge Graph (GKG), the technology Google uses to improve its search engine results. IE has also been useful in summary generation [1], fact verification [2], and other tasks requiring text data analysis.

Information extraction, like all other areas of natural language processing (NLP), has undergone many transformations. Although it was initially proposed as a document level task in the early days of the field, the systems of that time —rule-based ones— were not capable of solving the task efficiently. With the emergence of machine learning, annotated datasets were created, but due to the limitations

Doctoral Symposium on Natural Language Processing, 25 September 2025, Zaragoza, Spain.

mikel.zubillaga@ehu.eus (M. Zubillaga)

of the systems at that time, the task was simplified and defined at the sentence level [3]. The specific IE task that became the focus was event extraction, which consists of detecting the event trigger and extracting its corresponding arguments. In recent years, there have been two major paradigm shifts that have transformed the field: deep learning and the creation of large language models. Regarding information extraction, these new paradigms have brought great advances, both in low-data scenarios [4] and in low-resource languages [5]. However, these improvements have mostly occurred in sentence level IE, with document level IE receiving little attention. In this thesis, we will address the challenge of document level information extraction.

Deep learning. Currently, the most common approaches to IE are based on deep learning. Deep learning is distinguished by training neural networks with a large number of parameters. Although there are different neural network architectures, those based on Transformers [6] have dominated in language processing. Among other things, this type of architecture facilitated the emergence of pre-trained language models due to their efficient parallelization. These language models brought about the first paradigm shift in recent years: the so-called pre-train and fine-tune. This paradigm shift enabled the transfer of knowledge learned from large text corpora to tasks with little data, setting a new state of the art in all language processing tasks. Information extraction was not an exception [7, 8, 4]. Nevertheless, current approaches are limited to the sentence level, and it is not clear whether they will continue to be effective at the document level. In this thesis, the goal is to adapt these techniques to the document level.

Cross-lingual transfer. Thanks to deep learning and pre-trained language models, the ability known as cross-lingual transfer was discovered [9]. This ability refers to applying what is learned in one language to another language. In information extraction, this capability has been particularly important for implementing the task in languages other than English [10]. Although cross-lingual transfer enables the application of a system trained for English to another language, data is still needed in the target language to evaluate the system. On this subject, we created the first Basque event extraction dataset and examined the cross-lingual transfer capability on Basque data [5]. However, all the mentioned research has been done with small language models; therefore, this thesis will verify whether these techniques continue to be effective with large language models.

Large Language Models. Large Language Models (LLMs), and especially Large Generative Language Models —like ChatGPT— have also represented a paradigm shift in language processing. We have moved from the pre-train and fine-tune paradigm to pre-train and prompt [11]. In the last year, there have been many advances regarding IE, mainly due to the long contexts enabled by LLMs. For example, Sainz et al. [12] have shown that long contexts can be used to specify task guidelines, allowing language models to extract information without data. However, these advances still have difficulties generalizing beyond the sentence level. In this thesis, we will also try to address this challenge.

Document level IE. When information extraction was first defined, it was defined at the document level [13]. However, as mentioned earlier, the limitations of the technology at that time forced it to transform into a sentence level task. Today, some datasets go beyond sentences [14, 11]. Also for Basque, specifically the one developed by us [5]. However, these datasets work at the paragraph level rather than the document level, without dealing with the complexity of an entire document. The main document level dataset would be the granular task from BETTER [15], but the systems that have participated in it extract information sentence by sentence and then combine it. The aim of this thesis is to develop a system that directly extracts information given a document.

3. Description of the proposed research

This research proposal focuses on developing document level information extraction techniques using large language models (LLMs), with a particular emphasis on low-resource languages like Basque.

While language model-based techniques have achieved significant advances in IE, these advances have primarily been at the sentence level, historically due to the limited context available to supervised systems. However, today's large language models can process longer contexts (entire documents and beyond).

This thesis will investigate the development of models capable of extracting information at the document level using large language models, specifically in low-resource contexts where limited annotated data is available. The research will particularly focus on Basque and other low-resource languages.

The main task will be document level event extraction. The research will initially work in a supervised context, later limiting supervision to represent more realistic scenarios. Finally, it will evaluate performance in low-resource languages including Basque.

Goals:

1. **Find appropriate prompts for document level IE.** Language models receive the task description and the instance of the problem to be solved through text. However, it is not clear what is the best prompt for solving a task. This will be specifically investigated in this first task, as it will form the foundation for subsequent tasks. Additionally, a prompt that simultaneously solves all sub-tasks comprising the event extraction task will be sought, to avoid error propagation in task chains.

2. **Implement document level event extraction.** Unlike sentence level event extraction, the document level has its own problems. Prompt and evaluation are particularly important. In this task, these problems will be addressed by developing and evaluating an initial system. This system will be supervised with annotated data.

3. **Adapt low-resource data techniques to the document level.** As with sentence level —or even more- there is little annotated data at the document level. Therefore, it will be necessary to develop few-shot learning techniques. Since existing techniques are limited to the sentence level, their adaptation or, if necessary, the development of new techniques will be investigated.

4. **Implement (cross-lingual) knowledge transfer.** Although related to the previous point, the approach is different. In this case, the goal of this task is to find an answer to the question of how to adapt from a high-resource language with annotated data to a low-resource language. The techniques to be developed will utilize cross-lingual knowledge transfer necessary for solving event extraction. For this task, low-resource languages will be used to evaluate the system, with the focus of improving the state of the art for Basque.

5. **Implement document level IE for Basque.** As indicated in the fourth point, this thesis has a particular interest in Basque. In this objective, the general techniques developed in the fourth point will be adapted for Basque. And, if necessary, new annotated data will be created.

6. **Implement multi-document information extraction.** In this final objective, the knowledge acquired throughout the thesis will be used to make the leap from document level to multi-document event extraction. This objective aligns with the current —and near future- use of language models, being a task of great interest. However, it presents new challenges in itself. This work package will focus on the beginnings of this research line.

4. Methodology and the proposed experiments

Our methodology is based on the hypothesis that today's large language models can perform information extraction tasks —traditionally performed at the sentence level— at the document level or across multiple documents, by using their ability to work with long contexts. This extends to scenarios with few learning examples and for various languages. To explore this hypothesis, the general objectives mentioned in the previous section will be addressed according to a planned approach.

To prove that hypothesis an empirical method will be used: the proposed hypotheses will be implemented in a system and evaluated on publicly accessible datasets. This evaluation will compare our system with the state-of-the-art systems, and we will consider a hypothesis validated when we achieve statistically significant improvements in this comparison.

The objectives proposed for this thesis project are ambitious, and likely not all hypotheses will be confirmed. Therefore, through the empirical method, approaches will be tested one by one, focusing on the most promising ones while setting aside others. Once hypotheses are confirmed, descriptions of the systems and experiments built around these hypotheses will be submitted to the main conferences in our field (ICML, NeurIPS, ICLR, ACL, EACL, NAACL, EMNLP, all ICORE Class 1 - Core A or A*). At the end of the thesis, an article will be written for a high-impact-factor journal. The peer-review system will demonstrate that the research has been conducted according to international practices.

4.1. Research Tasks (RT) and Questions (RQ):

RT1: Prepare scenarios. Publicly available datasets will be used for evaluation to allow comparison with other state-of-the-art systems. Some document level IE datasets have already been identified and created, but it will be necessary to verify if new ones exist at the beginning of the thesis. At the same time, systems that have participated in these datasets will be analyzed, and the most important ones will be reimplemented to better understand the problem. The main research questions in this section will be:

RQ1.A) Whether the datasets are suitable for evaluating the benefits of the techniques developed in the project, and if not, whether it is possible to create such datasets.

RQ1.B) Conduct a quantitative and qualitative analysis to identify the advantages and weaknesses of state-of-the-art systems.

RT2: Develop a document level system. Large language models will be used to implement our first system to perform document level IE. This will require examining appropriate prompts and output representations for the task, comparing different language models, and proposing different learning techniques. The main research questions in this section will be:

RQ2.A) Identify appropriate prompts and output representations for solving the task. Since these prompts will depend on the model, we will also examine which model is most suitable for the task.

RQ2.B) Study which learning techniques are best for generalizing from sentence level to document level information extraction, while ensuring they function correctly at the sentence or segment level by evaluating them on standard datasets.

RT3: Adapt to scenarios with limited training data. Here we will examine how to implement techniques for acquiring, reusing, and adapting useful data in low-resource contexts. For example, techniques based on transfer learning and knowledge distillation will be explored. Additionally, we will also test cross-lingual techniques to extend what is learned in a high-resource language to low-resource languages. Therefore, the research questions associated with this task are:

RQ3.A) Investigate the weaknesses of the approach developed in RT2 in a scenario with limited learning data.

RQ3.B) Study what is the most appropriate way to perform transfer learning between document level datasets.

RQ3.C) Investigate whether and how sentence level datasets can be reused or reformulated for document level information extraction.

RQ3.D) Study which methods are most appropriate for executing cross-lingual transfer learning.

RT4: Implement multi-document information extraction. Finally, we will attempt to make the leap from document level to multi-document IE. This poses new challenges compared to document level information extraction. Addressing this ambitious research goal will require significant adaptations to the existing system. In this research task, we will try to answer the following questions:

RQ4.A) How can the developed system be adapted to work across multiple documents?

RQ4.B) What new challenges does this modality present?

RQ4.C) What are the weaknesses of the proposed approach in this modality and what are future research directions?

4.2. Schedule year by year

In the following paragraphs, we will analyze how we will organize this thesis year by year.

First Year - Foundations. First, research task RT1 will be initiated, trying to answer research questions RQ1.A and RQ1.B. Once the datasets are prepared, work will begin on research task RT2. For this, basic techniques will be developed and evaluated on these datasets. The following tasks are anticipated:

1.1) Collect the necessary datasets to carry out RQ1.A, and if it is necessary, create our own dataset.

1.2) To answer RQ1.B, state-of-the-art systems will be reimplemented and evaluated on the selected datasets.

1.3) Continuing with RQ1.B, a quantitative and qualitative analysis of the shortcomings of state-of-the-art systems will be conducted.

1.4) An article will be submitted to a major conference based on the answer to RQ1.B.

1.5) To carry out RQ2.A, an exploration of current large language models will be conducted to select the one that best fits the task.

1.6) Continuing with RQ2.A, different prompts will be designed to describe the task.

Second Year - Improving the State of the Art: In the second year, work will continue on research task RT2 with the aim of improving the state of the art. For this, what was learned in RQ1.B will be used. Additionally, work will begin on research task RT3, adapting the developed approach to scenarios with limited learning data. The following tasks are anticipated:

2.1) To answer RQ2.B, different approaches learned in research question RQ2.A will be developed and evaluated at both document and sentence/segment levels.

2.2) An article will be submitted to a major conference based on the answers to RQ2.A and RQ2.B.

2.3) Once the model from RQ2 is developed, it will be evaluated in scenarios with limited data to answer RQ3.A and its weaknesses will be analyzed. For this, the training data from existing datasets will be reduced, simulating data scarcity.

2.4) To address the weaknesses identified in RQ3.A, techniques known as knowledge transfer will be used. In RQ3.B, how this technique can be applied will be investigated.

2.5) Since document level data is scarce, RQ3.C will investigate how sentence level datasets can be reused for document level information extraction.

2.6) An article will be submitted to a major conference based on the answers to RQ3.A, B, and C.

Third Year - Implementing Cross-lingual Transfer: In the third year, work will continue on research task RT3. Specifically, techniques will be developed to make the approach function in low-resource languages. In particular, a system working in Basque will be developed. Finally, a research stay will be conducted at a university with experts in cross-lingual knowledge transfer and information extraction.

3.1) For RQ3.D, datasets in various languages will need to be collected first. If they don't exist, they will need to be created.

3.2) Continuing with RQ3.D, techniques based on cross-lingual transfer will be developed to extend an English-based approach to other languages.

3.3) Finally, what was learned in RQ3.D will be applied to Basque. For this, existing datasets will need to be adapted to the document level.

3.4) An article will be submitted to a major conference based on the answer to RQ3.D.

3.5) A research stay will be conducted.

Fourth Year - Completion and Writing: Taking the most interesting conclusions drawn from the exploration of research questions in previous years, work will be done on research task RT4. Since RT4 is an ambitious goal, work will be done primarily in the early stages of this topic. Finally, the thesis report will be written.

4.1) To answer question RQ4.A, necessary changes will be made to the developed approach.

4.2) To answer question RQ4.B, necessary datasets will be obtained or created.

4.3) Continuing with question RQ4.B, the difficulties and challenges shown by the adapted model will be analyzed using the aforementioned datasets.

4.4) What is learned from RQ4.B will be used to answer RQ4.C, defining post-thesis research directions.

4.5) An article will be submitted to a major conference with the conclusions drawn from RQ4.

4.6) The thesis report will be written.

5. Specific issues of research to be discussed

Our research focuses on document level Information Extraction (IE), which presents several methodological challenges. A primary limitation concerns dataset availability. Currently, we have identified two relevant datasets: MUC [13] and BETTER [15]. While both include multilingual components (with MUC's multilingual extension established by Gantt et al. [16]), neither encompasses Basque or other low-resource languages crucial to our research scope. Consequently, we will need to develop a custom Basque dataset for document level IE.

Evaluation methodology represents another significant challenge in document level IE research [17]. Our approach will implement the metrics framework proposed by Chen et al. [17], which addresses many of the unique challenges of this task.

Additionally, learning techniques for developing Large Language Models specialized in Information Extraction remain an open question. Our current investigations explore the application of reasoning-enhanced LLMs to improve IE performance. Specifically, we are examining Reinforcement Learning techniques to train these models to better handle complex document level information relationships.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT to: Grammar and spelling check, paraphrase, and translate. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

[1] Z. Zhang, H. Elfardy, M. Dreyer, K. Small, H. Ji, M. Bansal, Enhancing multi-document summarization with cross-document graph-based information extraction, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1696–1707. URL: https://aclanthology.org/2023.eacl-main.124/. doi:10.18653/v1/2023.eacl-main.124.

[2] J. Kim, K.-s. Choi, Unsupervised fact checking by counter-weighted positive and negative evidential paths in a knowledge graph, in: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 1677–1686. URL: https://aclanthology.org/2020.coling-main.147/. doi:10.18653/v1/2020.coling-main.147.

[3] C. Walker, S. Strassel, J. Medero, K. Maeda, Ace 2005 multilingual training corpus, Linguistic Data Consortium, Philadelphia 57 (2006) 45. URL: https://catalog.ldc.upenn.edu/LDC2006T06.

[4] O. Sainz, H. Qiu, O. Lopez de Lacalle, E. Agirre, B. Min, ZS4IE: A toolkit for zero-shot information extraction with simple verbalizations, in: H. Hajishirzi, Q. Ning, A. Sil (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations, Association for Computational Linguistics, Hybrid: Seattle, Washington + Online, 2022, pp. 27–38. URL: https://aclanthology.org/2022.naacl-demo.4/. doi:10.18653/v1/2022.naacl-demo.4.

[5] M. Zubillaga, O. Sainz, A. Estarrona, O. Lopez de Lacalle, E. Agirre, Event extraction in Basque: Typologically motivated cross-lingual transfer-learning analysis, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 6607–6621. URL: https://aclanthology.org/2024.lrec-main.586/.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[7] O. Sainz, O. Lopez de Lacalle, G. Labaka, A. Barrena, E. Agirre, Label verbalization and entailment for effective zero and few-shot relation extraction, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 1199–1212. URL: https://aclanthology.org/2021.emnlp-main.92/. doi:10.18653/v1/2021.emnlp-main.92.

[8] O. Sainz, I. Gonzalez-Dios, O. Lopez de Lacalle, B. Min, E. Agirre, Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Findings of the Association for Computational Linguistics: NAACL 2022,

Association for Computational Linguistics, Seattle, United States, 2022, pp. 2439–2455. URL: https://aclanthology.org/2022.findings-naacl.187/. doi:10.18653/v1/2022.findings-naacl.187.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.

[10] A. Subburathinam, D. Lu, H. Ji, J. May, S.-F. Chang, A. Sil, C. Voss, Cross-lingual structure transfer for relation and event extraction, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 313–325. URL: https://aclanthology.org/D19-1030/. doi:10.18653/v1/D19-1030.

[11] S. Li, H. Ji, J. Han, Document-level event argument extraction by conditional generation, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 894–908. URL: https://aclanthology.org/2021.naacl-main.69/. doi:10.18653/v1/2021.naacl-main.69.

[12] O. Sainz, I. García-Ferrero, R. Agerri, O. L. de Lacalle, G. Rigau, E. Agirre, GoLLIE: Annotation guidelines improve zero-shot information-extraction, in: The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=Y3wpuxd7u9.

[13] R. Grishman, B. Sundheim, Message Understanding Conference- 6: A brief history, in: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, 1996. URL: https://aclanthology.org/C96-1079/.

[14] S. Ebner, P. Xia, R. Culkin, K. Rawlins, B. Van Durme, Multi-sentence argument linking, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8057–8077. URL: https://aclanthology.org/2020.acl-main.718/. doi:10.18653/v1/2020.acl-main.718.

[15] T. Mckinnon, C. Rubino, The IARPA BETTER program abstract task four new semantically annotated corpora from IARPA's BETTER program, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3595–3600. URL: https://aclanthology.org/2022.lrec-1.384/.

[16] W. Gantt, S. Behzad, H. An, Y. Chen, A. White, B. Van Durme, M. Yarmohammadi, MultiMUC: Multilingual template filling on MUC-4, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 349–368. URL: https://aclanthology.org/2024.eacl-long.21/.

[17] Y. Chen, W. Gantt, W. Gu, T. Chen, A. White, B. Van Durme, Iterative document-level information extraction via imitation learning, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1858–1874. URL: https://aclanthology.org/2023.eacl-main.136/. doi:10.18653/v1/2023.eacl-main.136.