Cross-Lingual Transfer for Low-Resource Natural Language Processing

Transferencia crosslingüe para el Procesamiento del Lenguaje Natural

Iker García-Ferrero

Supervised by **German Rigau** and **Rodrigo Agerri**

HiTZ Zentroa - Ixa taldea
Euskal Herriko Unibertsitatea UPV/EHU

PhD Dissertation

February 12, 2025

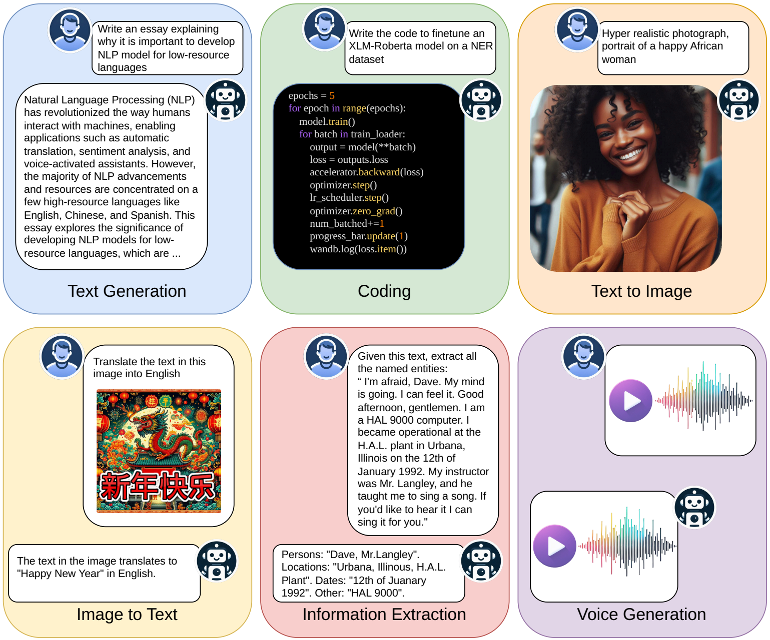
eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

OUTLINE

Introduction

Motivation



Transformer architecture (Vaswani et al., 2017) and neural networks have become an indispensable resource in NLP (Min et al., 2024).

- Trained on hundreds of terabytes of text data and billions of parameters.
- Can generate human-like text and have been applied in a wide range of applications.
- Hold the potential to bring significant societal changes (Bommasani et al., 2021).

Despite the remarkable progress in NLP, many challenges remain:

- LLMs require vast amounts of data and computational resources to achieve optimal performance (Hoffmann et al., 2022).
- Models consistently perform better on high-resource languages, especially English (Etxaniz et al., 2024). Their performance on low-resource languages is significantly lower (Ojo & Ogueji, 2023; Ojo et al., 2023).
- For the large majority of the approximately more than 7,000 languages spoken worldwide, training data is scarce or non-existent (Joshi et al., 2020).

Main Research Question

- *Develop cross-lingual transfer learning solutions to address the resource constraints faced by many languages, tasks, and domains.*

Cross-lingual transfer learning

Research area focused on creating models for low-resource languages by leveraging knowledge from high-resource languages.

Obama **PERSON** visited France **LOCATION** on Monday

We focus on **Sequence Labeling:**

- Assigning a label to each token in a given input sequence.
- Essential for: Information Extraction, Question Answering, and Sentiment Analysis, ...

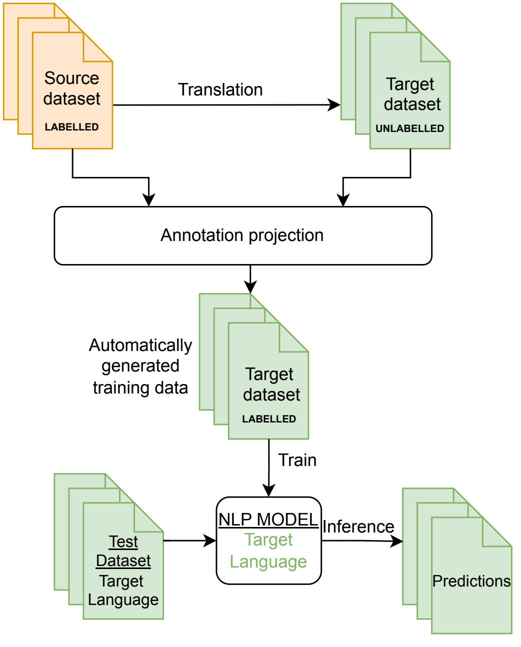
Introduction

Background

Data-Based Transfer

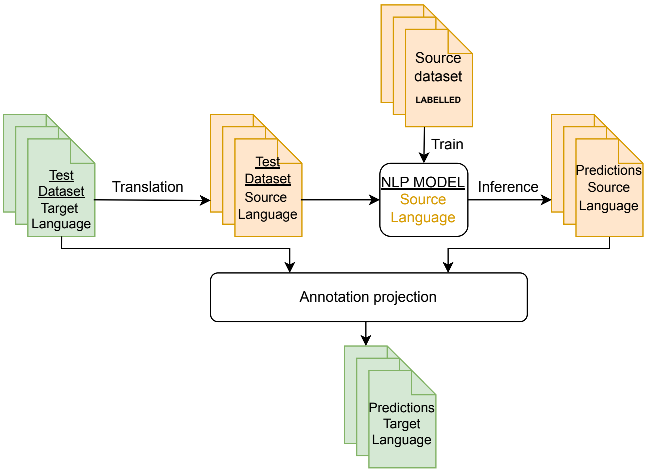
Use parallel data and/or Machine Translation to bridge the gap between languages in cross-lingual NLP tasks.

- The NLP model is trained and performs inference in the same language.
- There are two main approaches for data transfer: Translate-Train and Translate-Test.



Translate-Train

Automatically generate annotated data in languages where such data is scarce.



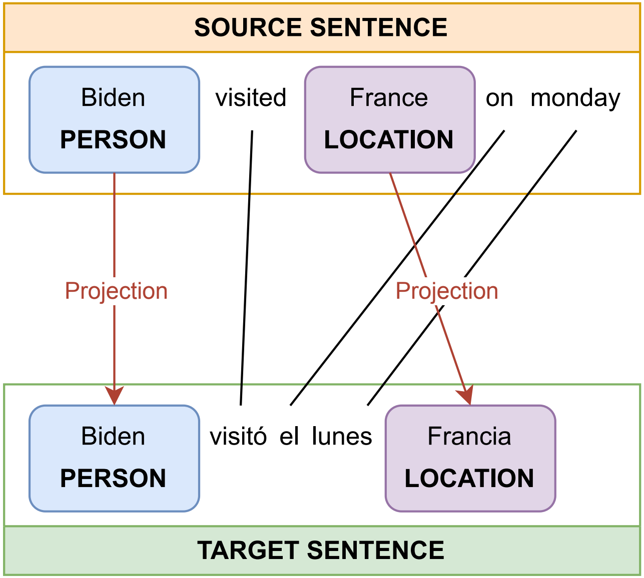
Translate-Test

Take advantage of the ability of the models to produce better results for high-resource languages such as English (Etxaniz et al., 2024):

Universidad del País Vasco Euskal Herriko Unibertsitatea
HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology

Annotation Projection

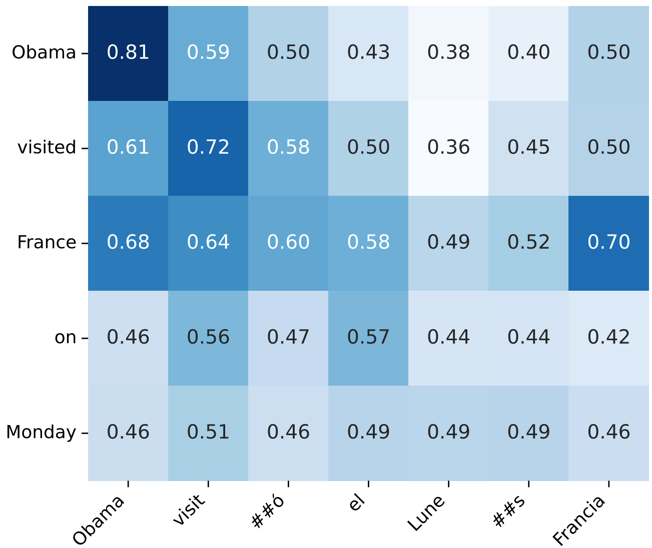
TASK	Example in source language	Translation	Label Projection Method
Text classification	Brazil won the World Cup Sports **TOPIC**	Brasil ganó la Copa del Mundo Sports **TOPIC**	None
Text Generation	Who is Freddie Mercury? Freddie Mercury was the lead voalist of the rock band Queen	¿Quién es Freddie Mercury? Freddie Mercury era el vocalista principal de la banda de rock Queen.	Translation
Sequence labeling	Obama **PERSON** visited France **LOCATION**	Obama **PERSON** visitó Francia **LOCATION**	Word Alignment



Annotation Projection with Word Alignments

Bidirectional graph between words in a parallel sentence.

- Statistical Machine Translation: Giza++ (Och & Ney, 2003), FastAlign (Dyer et al., 2013a), Eflomal (Östling & Tiedemann, 2016).
- Deep Learning Models: SimAlign (Jalili Sabet et al., 2020), AWESOME (Dou & Neubig, 2021).



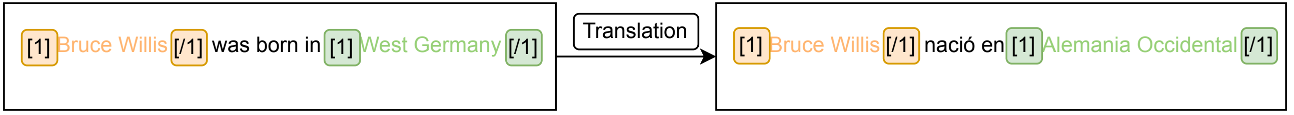
Deep-Learning based Word Alignments

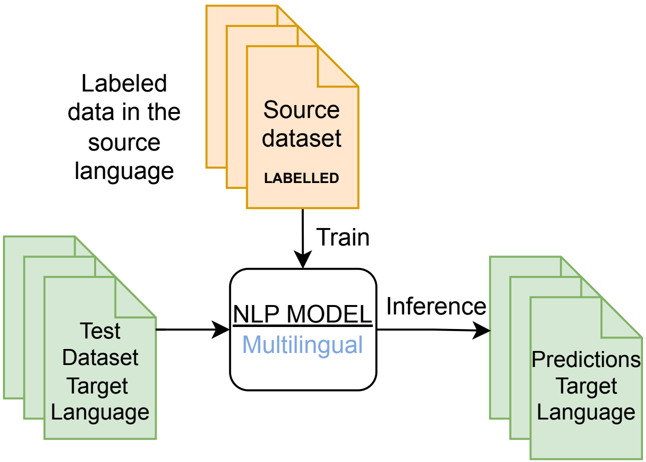
- SimAlign (Jalili Sabet et al., 2020): similarity of mBERT (Devlin et al., 2019) contextual embeddings.
- AWESOME: (Dou & Neubig, 2021) Unsupervised training on parallel data.

Other Annotation Projection methods

Replace word alignments in favor of directly using Machine Translation models.

- EasyProject (Chen et al., 2023): introduce markers in the source sentence. Translated together with the sentence.
- CODEC (Le et al., 2024): enhances this method by implementing a constrained decoding algorithm.





Model-based Transfer (Zero-shot)

Language models pre-trained on over 100 languages, such as BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), can be fine-tuned for a task in English and then used for inference in any of the languages included in the pre-training.

Data transfer vs Model transfer

Model and Data Transfer for Cross-Lingual Sequence Labelling in Zero-Resource Settings (EMNLP 2022)

Chapter Overview

- In-depth study of data transfer vs. model transfer for zero-shot cross-lingual sequence labeling.
- Previous studies were contradictory and did not incorporate the latest advancements in machine translation, word alignments, and sequence labeling models.
- Application of state-of-the-art machine translation, word alignments, and language models.
- **Objective**: Establish the conditions under which each approach—data transfer and zero-shot model-based cross-lingual transfer—outperforms the other.

Universidad del País Vasco / Euskal Herriko Unibertsitatea | HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology

Experimental Setup: Models

State-of-the-art models when this analysis was conducted:

- **Machine Translation: DeepL** [1], OpusMT (Tiedemann and Thottingal, 2020), mBART (mbart-large-50, Liu et al., 2020; Tang et al., 2020) and M2M100 (1.2B, Fan et al., 2021).
- **Word Alignments:** GIZA++ (Och & Ney, 2003), FastAlign (Dyer et al., 2013b), SimAlign (Jalili Sabet et al., 2020), **AWESOME** (Dou & Neubig, 2021).
- **Sequence Labeling Models:** mBERT (Devlin et al., 2019), XLM-RoBERTa (base and large) (Conneau et al., 2020).

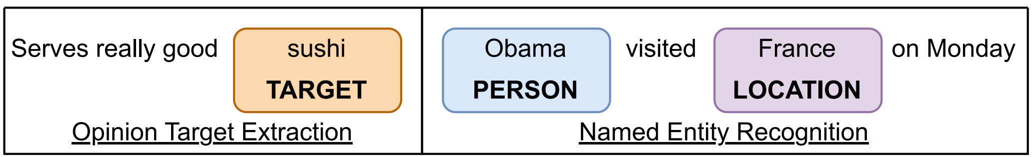
[1] https://www.deepl.com/es/translator

Data transfer vs Model transfer

Methodology

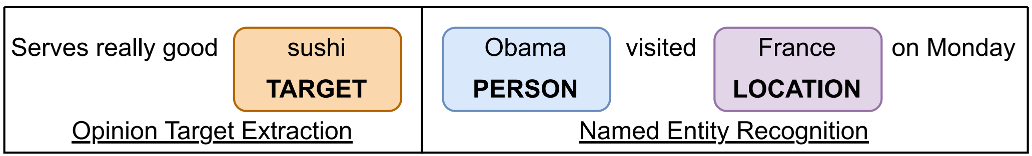
We focus on two Sequence Labelling tasks:

- Opinion Target Extraction (Pontiki et al., 2016): Given a review, the task is to detect the linguistic expression used to refer to the reviewed entity.
- Named Entity Recognition (Sang, 2002; Speranza, 2009): Given a text, the task is to detect named entities and classify them according to some pre-defined categories.



We assume the following scenario:

- We have English gold-labeled train and development data.
- Small amount of target language gold-labeled data is available for evaluation.
- No training data is available in the target language.



Data transfer vs Model transfer

Experimental results

mBERT			
Language	Zero-shot	Trans-Train	Trans-Test
English	-	-	-
Spanish	**68.4** $_{\pm 0.6}$	67.9 $_{\pm 0.8}$	62.2 $_{\pm 1.2}$
French	**62.7** $_{\pm 1.2}$	59.7 $_{\pm 1.2}$	57.6 $_{\pm 1.1}$
Dutch	61.7 $_{\pm 0.8}$	64.3 $_{\pm 1.5}$	**67.0** $_{\pm 0.8}$
Russian	53.8 $_{\pm 2.2}$	**62.9** $_{\pm 0.6}$	59.7 $_{\pm 0.4}$
Turkish	45.3 $_{\pm 4.0}$	**45.7** $_{\pm 2.3}$	35.5 $_{\pm 2.4}$
XLM-R base			
English	-	-	-
Spanish	**78.2** $_{\pm 0.4}$	72.5 $_{\pm 0.7}$	62.9 $_{\pm 0.9}$
French	**72.7** $_{\pm 0.3}$	64.7 $_{\pm 0.8}$	60.0 $_{\pm 0.6}$
Dutch	**75.5** $_{\pm 0.8}$	70.0 $_{\pm 1.6}$	71.0 $_{\pm 1.5}$
Russian	**74.9** $_{\pm 0.9}$	69.5 $_{\pm 0.3}$	62.2 $_{\pm 1.6}$
Turkish	58.1 $_{\pm 3.5}$	**58.9** $_{\pm 1.8}$	36.4 $_{\pm 1.8}$
XLM-R large			
English	-	-	-
Spanish	**79.3** $_{\pm 0.8}$	73.7 $_{\pm 1.1}$	64.0 $_{\pm 1.4}$
French	**74.6** $_{\pm 1.7}$	66.1 $_{\pm 0.6}$	60.7 $_{\pm 0.6}$
Dutch	**77.7** $_{\pm 1.9}$	74.0 $_{\pm 1.3}$	72.9 $_{\pm 1.8}$
Russian	**76.8** $_{\pm 1.3}$	69.3 $_{\pm 2.3}$	62.2 $_{\pm 1.3}$
Turkish	**62.4** $_{\pm 1.0}$	57.8 $_{\pm 2.4}$	33.7 $_{\pm 0.9}$

Opinion Target Extraction

- **mBERT:** Zero-shot better for Spanish and French. Data transfer superior for Dutch, Russian and Turkish.
- **XLM-R large:** Zero-shot superior for every language.
- **Translate-Train** is consistently superior to **Translate-Test**.

DATA TRANSFER VS MODEL TRANSFER

EXPERIMENTAL RESULTS

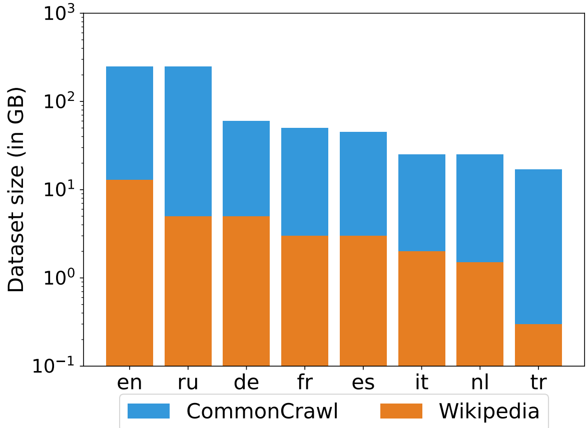
mBERT			
Language	Zero-shot	Trans-Train	Trans-Test
English	-	-	-
Spanish	**74.6**$\pm$0.4	69.5$\pm$0.4	70.8$\pm$0.6
German	**71.0**$\pm$0.9	70.1$\pm$0.3	70.6$\pm$0.5
Dutch	**78.5**$\pm$0.5	74.4$\pm$0.6	75.4$\pm$0.8
Italian	68.2$\pm$0.5	68.7$\pm$0.5	**70.7**$\pm$0.3
XLM-R base			
English	-	-	-
Spanish	**75.0**$\pm$0.4	70.1$\pm$0.6	72.5$\pm$0.2
German	67.9$\pm$0.5	**70.5**$\pm$0.5	70.1$\pm$0.8
Dutch	**78.1**$\pm$0.6	73.3$\pm$0.9	74.7$\pm$0.4
Italian	71.2$\pm$0.5	71.1$\pm$0.4	**71.7**$\pm$0.3
XLM-R large			
English	-	-	-
Spanish	**79.5**$\pm$1.0	70.9$\pm$0.6	74.0$\pm$0.5
German	**74.5**$\pm$0.7	73.7$\pm$0.5	72.9$\pm$0.3
Dutch	**82.3**$\pm$0.6	77.5$\pm$0.9	77.2$\pm$0.6
Italian	**76.0**$\pm$0.5	73.7$\pm$0.4	73.5$\pm$0.6

Named Entity Recognition

- **mBERT:** Zero-shot often outperforms data-based transfer methods.
- **XLM-R large:** Zero-shot consistently achieves the best results for all languages.
- **Translate-Test** is consistently superior to **Translate-Train**.

Data transfer vs Model transfer

Experimental results



Amount of data in GiB (log-scale) for the languages we use in our experiments in Wiki-100 (mBERT) and CC-100 (XLM-R.) from Conneau et al., 2020.

- mBERT's performance is better for languages topologically similar to English.
- XLM-R (both base and large) was trained with more data for Russian and Turkish than mBERT.
- Zero-shot performance relies on model proficiency in the target language and data domain.

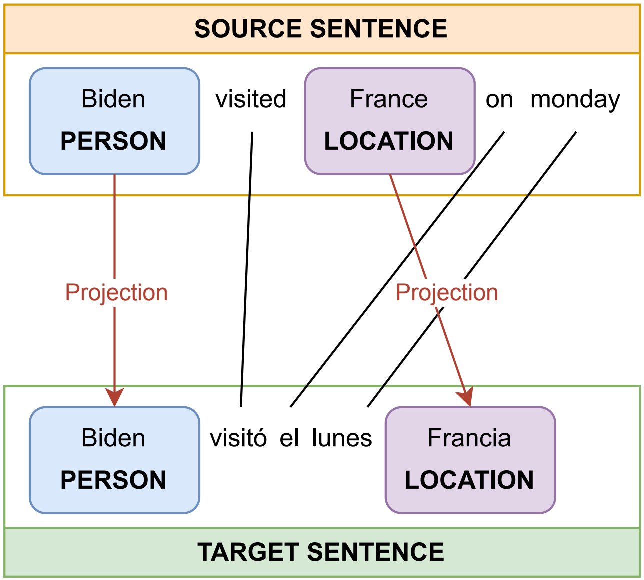
Conclusions:

- If you have a model proficient in both the source and target language $\rightarrow$ Model Transfer.
- Else $\rightarrow$ Data Transfer.

Improving Data Transfer

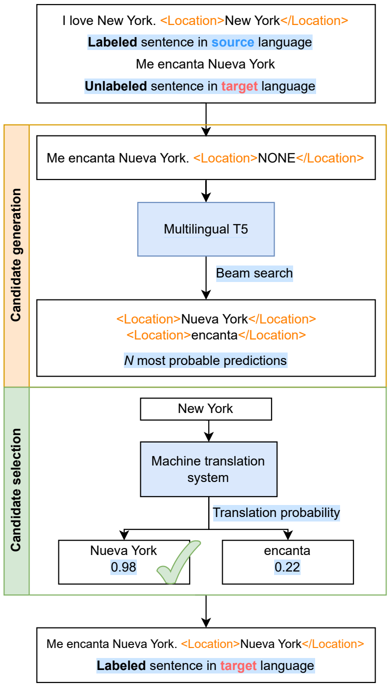
T-projection: High quality annotation projection for sequence labeling tasks. (EMNLP 2023)

Universidad del País Vasco · Euskal Herriko Unibertsitatea | HiTZ Hizkuntza Teknologiako Zentroa · Basque Center for Language Technology



Shortcomings of current protection models

- Word alignments often produce partial, incorrect or missing annotation projections.
- Based only on word co-occurrences or similarity between vector representations.

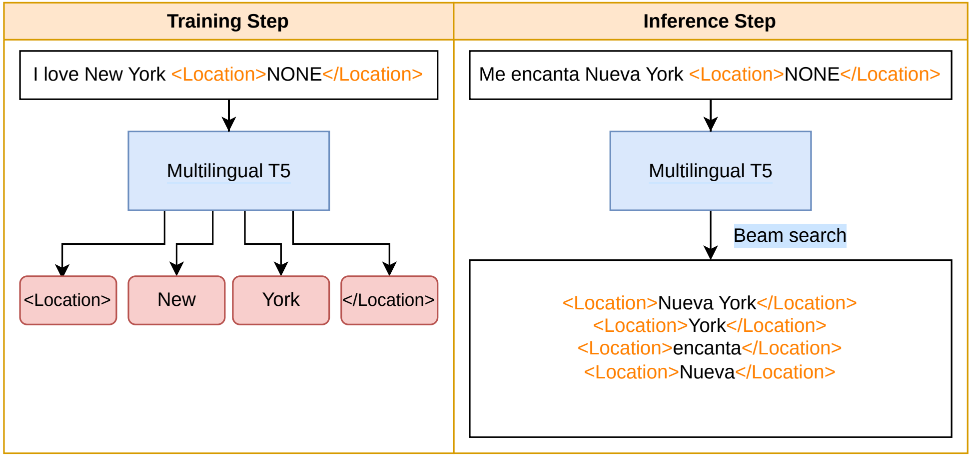


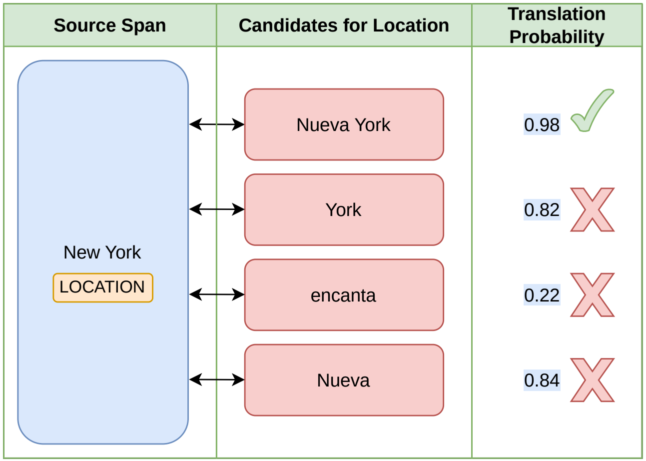
T-Projection

- We assume a set of source sentences with labeled spans. There is a parallel version of non-labeled sentences in a target language.
- Two main steps:
  - Candidate generation.
  - Candidate selection.

Candidate Generation

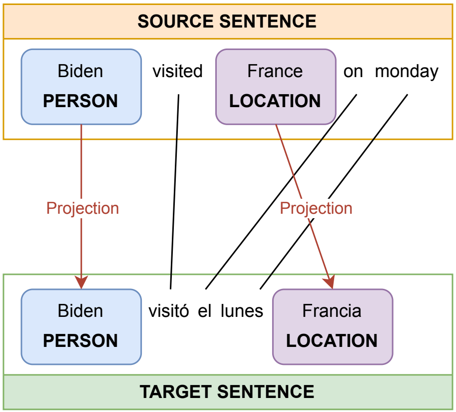
- Input: Text + Categories.
- Output: Replace *None* with the corresponding sequence.
- We generate 100 candidates using beam search.





Candidate selection

- Candidates not subsequence of the sentence are filtered out.
- Generated candidates are grouped by category.
- Candidates are ranked using translation probabilities from M2M100 (Fan et al., 2021) or NLLB200 (Costa-jussà et al., 2022).



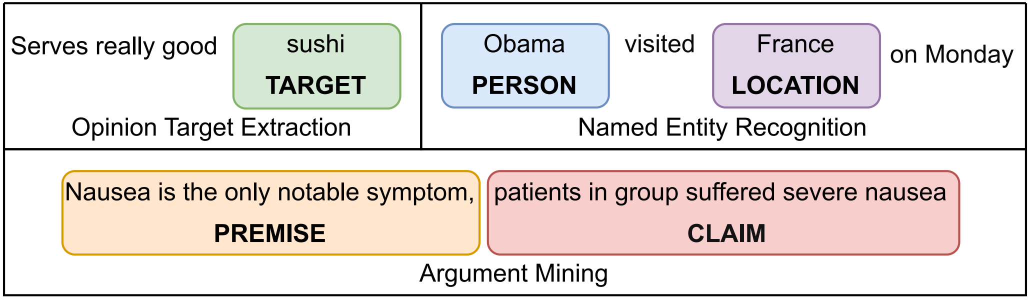
Baselines

- **Word alignment systems** (Giza++, FastAlign, SimAlign, AWESOME).
- **XLM-RoBERTa**: Train with the English labeled data, annotate the parallel target sentences (B. Li et al., 2021).
- **Translation based projection**: Translate-Match, EasyProject, CODEC.

Intrinsic Evaluation: Datasets

Manually projected datasets:

- **Opinion Target Extraction (OTE)** SemEval 2016 English datasets (Restaurant domain), manual label projections in Spanish, French, and Russian.
- **Named Entity Recognition (NER)**: parallel data in English, Spanish, German, and Italian (Europarl). For extrinsic eval: MasakhaNER 2.0.
- **Argument Mining (AM)**: AbstRCT English dataset (Mayer et al., 2020), Spanish parallel version.



Universidad del País Vasco Euskal Herriko Unibertsitatea | HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology

Intrinsic Evaluation: Annotation Projection Quality

	OTE			NER			AM	Avg
	ES	FR	RU	ES	DE	IT	ES	
Giza++ (Och and Ney, 2003)	77.0	73.3	72.4	73.3	75.3	68.4	86.6	77.7
FastAlign (Dyer et al., 2013b)	75.0	72.9	76.9	70.2	77.0	67.0	85.7	77.4
SimAlign (Jalili Sabet et al., 2020)	86.7	86.3	87.7	85.4	87.4	81.3	84.1	85.3
AWESOME (Dou and Neubig, 2021)	91.5	91.1	93.7	87.3	90.7	83.1	54.8	78.0
XLM-RoBERTa-xl (Conneau et al., 2020)	80.2	76.2	74.5	73.9	68.3	73.9	66.5	71.8
Span Translation	66.5	46.3	58.7	68.8	63.5	69.2	21.6	48.7
T-Projection	**95.1**	**92.3**	**95.0**	**93.6**	**94.0**	**87.2**	**96.0**	**93.9**

Table. F1 scores for annotation projection in the OTE, NER and Argument Mining tasks.

Experimental Setup for the Extrinsic Evaluation

- The English CoNLL data set is translated into the 8 African languages using NLLB200.
- We project the English gold labels into the automatically translated parallel data.
- We train XLM-R-large with the African languages' silver data.
- We evaluated XLM-R-large on a gold-labeled test dataset in the 8 African languages.

EXTRINSIC EVALUATION

Universidad del País Vasco Euskal Herriko Unibertsitatea

HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology

Language	No. of Speakers	Language family	Zero Shot	AWESOME +English	EasyProject +English	CODEC	T-Projection	T-Projection +English
Hausa	63M	Afro-Asiatic /Chadic	71.7	**72.7**	72.2	72.4	**72.7**	72.0
Igbo	27M	NC / Volta-Niger	59.3	63.5	65.6	70.9	71.4	**71.6**
Chichewa	14M	English-Creole	**79.5**	75.1	75.3	76.8	77.2	77.8
chiShona	12M	NC / Bantu	35.2	69.5	55.9	72.4	**74.9**	74.3
Kiswahili	98M	NC / Bantu	**87.7**	82.4	83.6	83.1	84.5	84.1
isiXhosa	9M	NC / Bantu	24.0	61.7	71.1	70.4	**72.3**	71.7
Yoruba	42M	NC / Volta-Niger	36.0	38.1	36.8	41.4	**42.7**	42.1
isiZulu	27M	NC / Bantu	43.9	68.9	73.0	**74.8**	66.7	64.9
AVG			54.7	66.5	66.7	**70.3**	**70.3**	69.8

Table. F1 scores on MasakhaNER2.0 for mDebertaV3 trained with projected annotations from different systems. "+EN" denotes concatenation of the automatically generated target language dataset with the source English dataset.

- T-Projection outperforms current state-of-the-art label projection systems in both intrinsic and extrinsic evaluations by a wide margin.
- Data-based transfer approaches such as T-Projection can be highly effective for performing NLP tasks in low-resource languages.

Improving Model Transfer

Paper submitted for review (2025)

Motivation

- Model transfer with high-capacity models is effective for cross-lingual tasks.
- Text-to-text Large Language Models (LLMs) are the most powerful models.

Universidad del País Vasco Euskal Herriko Unibertsitatea
HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology

LLMs vs Encoder Models

- Encoder-only models such as XLM-RoBERTa have around 561M parameters trained on 295B tokens.
- Text-to-text LLMs such as T5, LLaMA and GPT-4 have significantly more parameters and were trained on much larger datasets.

	XLM-RoBERTa Conneau et al., 2020	XLM-RoBERTa-xxl Goyal et al., 2021	mT5 Xue et al., 2021	Llama2 Touvron et al., 2023	Gemma2 Mesnard et al., 2024	LLama3 AI@Meta, 2024
Parameters	560M	10.7B	11.3B	70B	27B	405B
Train Tokens	296B	296B	1T	2T	8T	17T

Table. Size and training data of some relevant open source models.

Universidad del País Vasco Euskal Herriko Unibertsitatea
HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology

LLMs vs Encoder Models

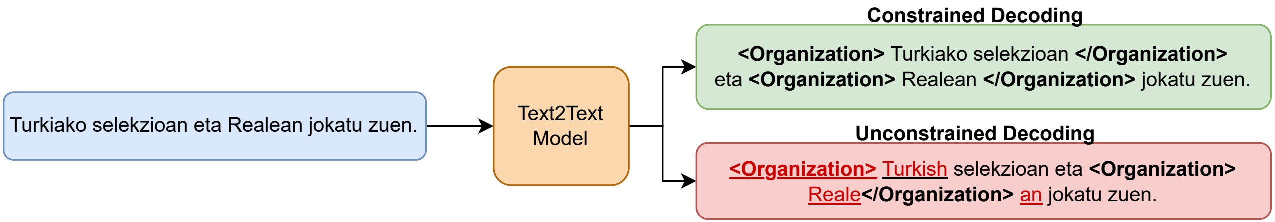
- Text-to-Text LLM do not work out-of-the-box for cross-lingual sequence labelling.

Model	**Size**	**amh**	**bam**	**bbj**	**ewe**	**hau**	**ibo**	**kin**	**lug**	**luo**	**mos**	**nya**	**pcm**	**sna**	**swa**	**tsn**	**twi**	**wol**	**xho**	**yor**	**zul**
`Fine-tune: SotA`																					
AfroXLMR-large	550M	**78.0**	**79.0**	**90.3**	75.2	**85.4**	**88.9**	**86.8**	**88.9**	**75.3**	**73.5**	**92.4**	**90.0**	**96.1**	**92.7**	**88.9**	**79.2**	**83.8**	**89.2**	**67.9**	**90.6**
`Prompting of LLMs`																					
GPT-4	-	28.5	52.7	50.3	**75.6**	64.9	56.0	55.1	73.3	49.8	60.2	63.6	64.7	33.4	71.5	64.6	58.6	67.9	28.4	58.3	34.9
AYA	-	14.1	7.1	20.0	26.5	34.5	28.2	30.8	16.3	12.7	34.4	21.7	27.4	13.4	35.6	29.4	18.9	14.5	4.2	17.5	11.4
mT0	13B	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
mT0-MT	13B	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaMa 2	13B	0.0	13.8	12.3	25.1	22.1	22.0	23.1	27.5	19.0	11.0	20.0	27.5	11.3	25.8	26.2	20.7	16.0	8.1	15.1	9.0

Table. Comparison of F1-score of various LLMs with that of the current state of the art result in Masakhaner 2.0. Table reproduced from Ojo and Ogueji, 2023.

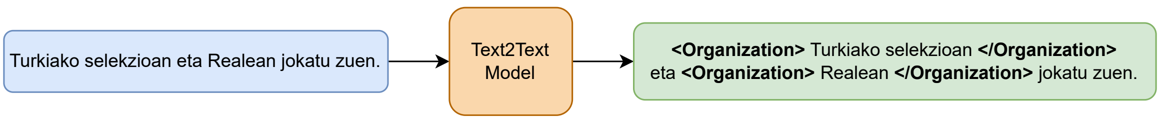
Challenges with LLMs in Zero-Shot Sequence Labeling

- Text-to-text models are designed for free-form text generation.
- Models do not strictly adhere to the expected output structure (e.g., tags).
- Outputs often mix source and target languages.
- Outputs can hallucinate non-existing spans.



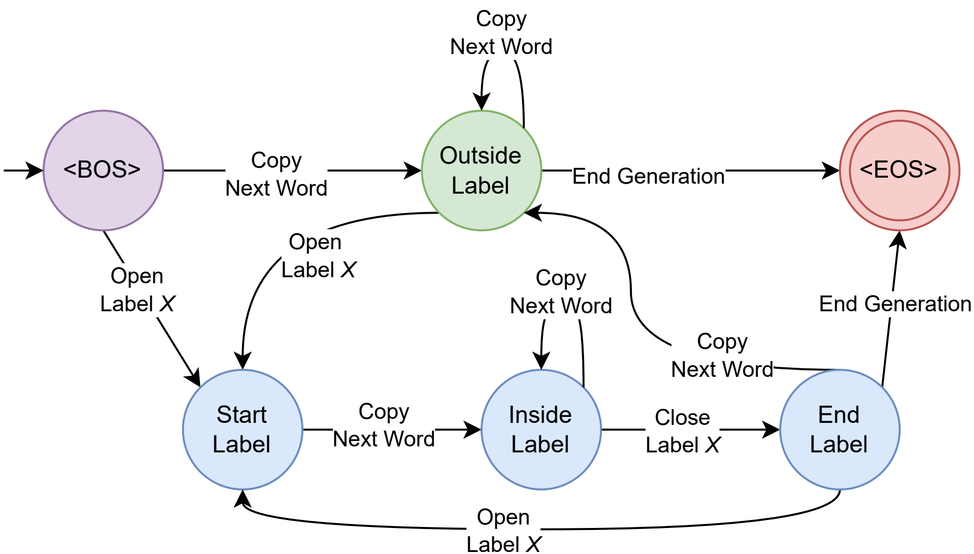
Input-Output Representation

- The expected output is the same sentence annotated with HTML-style tags.
- Other task representations can be used with our method.



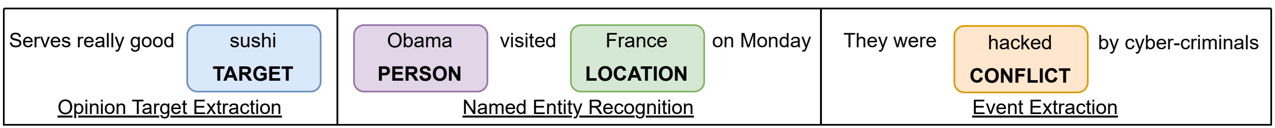
Finite State Automaton

Our Constrained Decoding Algorithm is defined as a Finite State Automaton.



Information Extraction Tasks

- **Named Entity Recognition (NER)**: MasakhaNER 2.0 (20 African languages), trained with English CoNLL03.
- **Opinion Target Extraction (OTE)**: SemEval 2016 train with English dataset, test in Spanish, French, Dutch, Russian, and Turkish.
- **Event Extraction (EE)**: ACE05 (Walker et al., 2006) trained in English, tested in Chinese.



Universidad del País Vasco Euskal Herriko Unibertsitatea

HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology

Language Models and Baselines

- **Baselines**:
  - Unconstrained decoding (**Base**).
  - Encoder-only models: mDeBERTa-v3 (He et al., 2021), GLOT500 (Imani et al., 2023), XLM-RoBERTa (Conneau et al., 2020) and afro-xlmr-large (Alabi et al., 2022).
- **Text-to-text Models**:
  - Encoder-decoder: mT0-XL (Muennighoff et al., 2023), mT5 (Xue et al., 2021), Aya-101 (Üstün et al., 2024).
  - Decoder-only: Qwen2 (Yang et al., 2024), Gemma (Team et al., 2024), LLaMA-3 (AI@Meta, 2024), Aya-23 (Aryabumi et al., 2024), and Yi 1.5 (AI et al., 2024).

Evaluation Metrics:

Standard F1-score metric for Sequence Labeling. Model output converted to IOB2 format. Evaluation performed with the seqeval library.

Model	Unconstrained	Constrained	Delta
mT5-xl	62.4	**65.7**	+3.3
mT0-xl	59.8	**65.7**	+5.9
aya-101	58.4	60.1	+1.7
Qwen2-7B-Instruct	39.7	42.0	+2.3
gemma-1.1-7b-it	46.8	49.0	+2.2
Llama-3-8B-Instruct	51.2	52.7	+1.6
aya-23-8B	51.6	52.6	+0.9
Yi-1.5-9B-Chat	52.8	57.1	+4.3
GLOT500	59.6		
mDeBERTa-v3	55.1		
afro-xlmr-large	58.7		

Table. Average F1 scores in the MasakhaNER dataset.

	mT0-xl		GLOT	mDeBERTa
Lang	Base	Cons	500	V3
English	82.6	84.8	82.6	83.6
Spanish	77.8	79.4	69.4	78.0
French	74.1	76.6	65.8	76.9
Dutch	74.1	77.1	66.5	77.3
Russian	71.1	75.7	69.2	76.5
Turkish	56.8	57.7	50.4	56.4
Average	70.8	73.3	64.3	73.0

	mT0-xl		GLOT	mDeBERTa
Lang	Base	Cons	500	V3
$\text{English}_{\text{Entity}}$	95.5	95.5	94.5	95.3
$\text{Chinese}_{\text{Entity}}$	70.1	73.3	34.1	54.2
$\text{English}_{\text{Trigger}}$	78.9	78.9	74.1	78.0
$\text{Chinese}_{\text{Trigger}}$	49.6	52.1	0.0	30.5

Conclusions

- Constrained Beam Search enables the use of multilingual text-to-text LLMs for cross-lingual model transfer.
- For the first time, we achieve better results than encoder-only models.

ODESIA LEADERBOARD
EVALUACIÓN COMPARADA DE MODELOS DE LENGUAJE ESPAÑOL/INGLÉS

Sistema	Team	Media aritmética	EXIST 2022: Sexism detection	EXIST 2022: Sexism categorisation	DIPROMATS 2023: Propaganda identification	DIPROMATS 2023: Coarse propaganda characterization	DIPROMATS 2023: Fine-grained propaganda characterization
Qwen2.5-14B-Instruct	ixa_taldea	**0.6306**	0.8027	0.6065	0.8360	0.5530	0.4931
xlm_roberta_cpt_en_es v2	BSC_models	**0.6237**	0.7816	0.6004	0.8166	0.5756	0.4837
Llama_3.1-8B-Instruct 0 shot no BIO v4	GPLSI	**0.6012**	0.7989	0.6203	0.8274	0.5379	0.4383
Llama3.1-8B-NoPrompt	ODESIA	**0.5886**	0.7490	0.5765	0.8054	0.5572	0.4521
XLM-RoBERTa-large-v3	UMUTeam	**0.5462**	0.7452	0.5540	0.8224	0.5425	0.4581
RigoBERTa	IIC	**0.5264**	0.7490	0.5957	0.8133	0.5594	0.4670
DeepSeek_Llama3.1	UDA-LIDI	**0.5163**	0.7586	0.5077	0.7534	0.4525	0.3687

Medical mT5

Medical mT5: An Open-Source Multilingual Text-to-Text LLM for The Medical Domain. (LREC-COLING 2024)

State-of-the-art in the Medical domain models at the start of this project.

Model	Reference	# Param	Text2Text	Multilingual
XLM-RoBERTa	Conneau et al. 2019	250M–12B	No	Yes
mDeBERTa-v3	He et al. 2020	86M	No	Yes
BioBERT	Lee et al. 2019	110M	No	No
PubMedBERT	Gu et al. 2020	110M	No	No
SciFive	Phan et al. 2021	220M–770M	Yes	No
BSC-BIO	Carrino et al. 2022	125M	No	No
BioLinkBERT	Yasunaga et al. 2022	110M–340M	No	No
BioT5X	Phan et al. 2022	110M–340M	Yes	No
BioGPT	Luo et al. 2022	347M	Yes	No
BioMedLM	Venigalla et al. 2022	2.7B	Yes	No
Med-PaLM	Singhal et al. 2022	540B	Yes	No
EriBERTa	To be published	–	No	Yes
Our Medical mT5	–	738M–3B	Yes	Yes

What do we need to build a text-to-text model for the Medical Domain?

- Compiling a Multilingual Corpus for the Medical Domain.
- Train a Multilingual model.
- Develop Multilingual evaluation benchmarks.
- Evaluate the model.

Compiling a Multilingual Corpus

Language	Source	Words
English	ClinicalTrials	127.4M
	EMEA	12M
	PubMed	968.4M
	Total	**1.1B**
Spanish	EMEA	13.6M
	PubMed	8.4M
	Medical Crawler	918M
	SPACC	350K
	UFAL	10.5M
	WikiMed	5.2M
	Total	**960M**
French	PubMed	1.4M
	Science Direct	15.2M
	Wikipedia - Médecine	5M
	EDP	48K
	Google Patents	654M
	Total	**676M**
Italian	Medical Commoncrawl - IT	67M
	Drug instructions	30.5M
	Wikipedia - Medicina	13.3M
	E3C Corpus - IT	11.6M
	Medicine descriptions	6.3M
	Medical theses	5.8M
	Medical websites	4M
	PubMed	2.3M
	Supplement description	1.3M
	Medical notes	975K
	Pathologies	157K
	Medical test simulations	26K
	Clinical cases	20K
	Total	**143M**
Total		**3.02B**

Multilingual Medical Corpus Overview

- 3 Billion words in English, Spanish, French, and Italian.
- Diverse public data sources.
- Focus on medical texts.

What do we need to build a text-to-text model for the Medical Domain?

- Compiling a Multilingual Corpus for the Medical Domain.
- Train a Multilingual model.
- Develop Multilingual evaluation benchmarks.
- Evaluate the model.

Pre-training Details

- Flax implementation, Hugging Face Transformers.

	Medical-mT5-large	Medical-mT5-xl
Param. no.	738M	3B
Sequence Lenght	1024	480
Token/step	65536	30720
Epochs	1	1
Total Tokens	4.5B	4.5B
Optimizer	Adafactor	Adafactor
LR	0.001	0.001
Scheduler	Constant	Constant
Hardware	4xA100	4xA100
Time (h)	10.5	20.5
CO_2eq (kg)	2.9	5.6

Table. Pre-Training settings for Medical mT5.

What do we need to build a text-to-text model for the Medical Domain?

- Compiling a Multilingual Corpus for the Medical Domain.
- Train a Multilingual model.
- Develop Multilingual evaluation benchmarks.
- Evaluate the model.

Multilingual Benchmark Challenges

- Lack of multilingual benchmarks in the medical domain.
- Existing datasets often English-centric.

Data Transfer

- Leveraging data-transfer techniques.
- Generate French, Spanish, Italian benchmarks from English data.
- Focus on: Argument Mining, Question Answering.

Universidad del País Vasco Euskal Herriko Unibertsitatea | HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology

Argument Mining: Data Generation

- Same method as for Spanish in Yeginbergen et al., 2024.
- English data -> Machine Translated into other languages
- Label Projection
- Manual Review.

Question Answering

- BioASQ-6B English dataset.
- Question + Context -> Generate Answer.

Data Generation

- Machine Translate Questions and Answers.
- Manual review of translations.

Evaluation Datasets

- Sequence Labeling: NER (E3C, DIANN), Argument Mining (AbstRCT).
- Generative Question Answering: BioASQ.

Representation	Task	Dataset	Languages	Entity Type
Sequence Labelling	Named Entity Recognition	NCBI-Disease, Dogan et al., 2014	EN	Disease
		BC5CDR Disease, J. Li et al., 2016	EN	Disease
		BC5CDR Chemical, J. Li et al., 2016	EN	Chemical
		DIANN, Fabregat et al., 2018	EN, ES	Disability
		E3C, Magnini et al., 2021	EN, ES, FR, IT	Clinical Entity
		PharmaCoNER, Gonzalez-Agirre et al., 2019	ES	Pharmacological
	Argument Mining	AbstRCT, Mayer et al., 2021	EN, ES, FR, IT	Claims and Premises
Generative Question Answering	Question Answering	BioASQ 6B, Tsatsaronis et al., 2015	EN, ES, FR, IT	Biomedical QA

Experimental Setup

Text-to-Text Conversion

- Sequence Labeling: HTML-style tags.
- Constrained decoding.
- Question Answering: Question and snippets as context -> Answer generation

Patient with dilated cardiomyopathy .

↓

Patient with dilated **<Disease>** cardiomyopathy **</Disease>**.

Experimental Setup

Question

question: Describe mechanism of action of Napabucasin. **context**: Napabucasin (BBI608) is an orally administered small [...]. The STAT3 transcription factor inhibitor, BBI608 [..]

snippet 1

snippet 2

Napabucasin (BBI608) is an orally administered small molecule that blocks stem cell activity in cancer cells by targeting the signal transducer and activator of transcription 3 (STAT3) pathway

What do we need to build a text-to-text model for the Medical Domain?

- Compiling a Multilingual Corpus for the Medical Domain.
- Train a Multilingual model.
- Develop Multilingual evaluation benchmarks.
- Evaluate the model.

Experimental Results

Sequence labeling Tasks

Lang	Dataset	$mT5_{large}$	$mT5_{XL}$	SciFive	$FlanT5_{large}$	$FlanT5_{XL}$	$mDeBERTa_{V3\ base}$	BioBERT	$MedMT5_{large}$	$MedMT5_{XL}$
EN	NCBI-Disease	85.1	87.7	**89.4**	88.6	89.3	85.7	87.4	89.1	87.2
EN	BC5CDR Disease	78.5	81.4	85.4	85.0	**85.8**	82.5	84.3	84.4	82.4
EN	BC5CDR Chemical	89.1	90.8	**93.3**	92.0	92.9	91.1	92.9	92.8	91.3
EN	DIANN	70.1	77.8	71.9	74.4	74.2	**80.3**	79.0	74.8	77.6
ES	DIANN	72.4	74.9	70.5	70.7	70.9	**78.3**	70.2	74.9	74.8
EN	E3C	54.3	60.1	62.8	**64.2**	63.1	58.2	58.6	59.4	57.9
ES	E3C	61.6	71.7	62.7	64.4	67.1	65.9	57.4	**72.2**	69.5
FR	E3C	55.6	64.9	61.7	65.2	64.3	62.0	53.3	65.2	**65.8**
IT	E3C	61.8	63.8	59.6	61.9	65.1	63.9	52.1	**67.5**	65.9
ES	PharmaCoNER	86.3	90.6	87.5	88.5	89.1	89.4	88.6	**90.8**	90.1
EN	Neoplasm	70.4	71.1	74.4	**74.3**	73.4	64.5	67.5	73.9	73.2
EN	Glaucoma	70.7	75.1	77.1	**78.4**	78.0	71.2	74.8	76.2	76.4
EN	Mixed	68.5	73.0	73.4	73.2	**74.5**	63.4	69.6	72.2	72.0
ES	Neoplasm	69.0	56.1	71.4	72.5	**73.9**	63.0	57.1	72.1	71.8
ES	Glaucoma	69.3	70.7	73.9	73.8	75.2	68.6	64.5	**77.1**	75.5
ES	Mixed	68.4	66.2	69.2	69.3	71.6	61.3	58.9	**72.4**	71.4
FR	Neoplasm	70.5	66.6	**74.0**	72.4	73.7	63.9	59.0	72.9	71.2
FR	Glaucoma	71.1	69.2	77.8	74.8	77.2	60.3	65.6	**79.5**	75.8
FR	Mixed	68.3	65.4	72.0	70.9	**74.3**	64.1	61.3	73.3	69.7
IT	Neoplasm	68.1	69.9	70.1	70.9	72.0	64.4	54.8	71.2	**73.1**
IT	Glaucoma	69.2	71.5	73.7	74.0	75.9	74.7	65.8	75.7	**78.7**
IT	Mixed	66.3	67.7	67.4	69.9	70.0	61.3	57.4	70.6	**71.9**
AVERAGE		70.2	72.1	73.6	74.1	75.1	69.9	67.3	**75.4**	74.7
AVERAGE ES, FR, IT		68.4	69.2	70.8	71.4	72.9	67.2	61.9	**74.0**	73.2

Single Task Supervised Results

- Medical-mT5 outperforms mT5, showing benefits of medical corpus.
- SciFive and Flan-T5 are superior on English (English specific).
- Medical-mT5 better for FR, ES, IT.

Evaluation Setup

- Manual evaluation by medical professionals.
- Ranked generated answers.
- Focus on Medical-mT5, mT5, FlanT5, SciFive

Universidad del País Vasco Euskal Herriko Unibertsitatea
HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology

Observations

- Models generated similar answers (difficulty to distinguish).
- All models failed on the same difficult questions.
- Low inter-annotator agreement (IAA)

Ideal Answer	Which type of urinary incontinence is diagnosed with the Q tip test? Stress urinary incontinence is diagnosed with the Q tip test. The test evaluates urethral mobility.
mT5-large	Q-tip test is used for urodynamic assessment of patients with severe urinary incontinence.
FlanT5-large	The Q tip test is used for assessing urethral mobility.
SciFive	Q tip test is used for evaluation of stress urinary incontinence.
MedMT5-large	Q-tip test is used for evaluation of stress urinary incontinence.
Ideal Answer	Which are the main manifestations of Ohdo syndrome? Severe ID, absent or deficient language, skeletal manifestations including bilateral patella dislocations.
mT5-large	Skeletal manifestations in Ohdo syndrome are a case with bilateral patella dislocations where surgical intervention has been indicated.
FlanT5-large	The main manifestations of Ohdo syndrome are: 1) severe ID, 2) absent or deficient language and 3) milder, clinical manifestation in heterozygotes.
SciFive	Ohdo syndrome is characterized by severe ID, absent or deficient language and, milder, clinical manifestation in heterozygotes.
MedMT5-large	The main manifestations of Ohdo syndrome are: 1) absent or deficient language and 2) mildder clinical manifestation in heterozygotes.

Summary

- Introduced Medical mT5, open-source multilingual medical LLM.
- New multilingual corpus (3B words).
- Evaluation benchmarks (AM, QA) generated.
- Superior performance in multi-task, zero-shot settings.
- Challenges in evaluating generative tasks.

Conclusions and Future Work

In this thesis, we have made the following contributions:

- Model vs. Data cross-lingual transfer evaluation.
- Improve data transfer: T-Projection.
- Improve model Transfer: Constrained decoding.
- Medical mT5 Framework

Software

NoticIA
A LLM finetuning and LLM evaluation library for the NoticIA dataset. The dataset consisting of 850 Spanish news articles featuring prominent clickbait headlines, each paired with high-quality, single-sentence generative summarizations written by humans.
- GitHub Repository

Clickbait Fighter
An AI that generates one-sentence summaries of sensational and clickbait news articles, which is used daily by Spanish users. I crafted the training dataset by hand. I trained the model on 8 A100 GPUs, and the demo runs on the OmegaAI cloud, utilizing vLLM and Ray. User feedback is used to continuously improve the model.
- Link to the app

GoLLIE
We present GoLLIE, a Large Language Model trained to follow annotation guidelines. GoLLIE outperforms previous approaches on zero-shot Information Extraction and allows the user to perform inferences with annotation schemas defined on the fly. Different from previous approaches, GoLLIE is able to follow detailed definitions and does not only rely on the knowledge already encoded in the LLM.
- GitHub Repository

T-Projection
T-Projection is a method to perform high-quality Annotation Projection of Sequence Labeling datasets. The code is built on top of 🤗HuggingFace's Transformers and 🤗HuggingFace's Accelerate library.
- GitHub Repository

Sequence Labeling with LLMs
Sequence Labelling with LLMs is a library code for performing Sequence Labelling with Language Models (LLMs) as a Text2Text constrained generation task. The code is built on top of 🤗 HuggingFace's Transformers and 🤗HuggingFace's Accelerate library.
- GitHub Repository

LM Contamination Index
Contaminated
The LM Contamination Index is a manually created database of contamination evidences for LMs. Please
- Web Page

Datasets

updated less than a minute ago

This is a list of LLMs I have helped develop.

This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models
Paper • 2310.15941 • Published Oct 24, 2023 • ▲ 6

HiTZ/This-is-not-a-dataset
Viewer • Updated Feb 23, 2024 • 381k • 191 • 6

HiTZ/Multilingual-Opinion-Target-Extraction
Viewer • Updated Nov 22, 2023 • 12.7k • 154 • 1

HiTZ/Multilingual-Medical-Corpus
Viewer • Updated Apr 12, 2024 • 67.4M • 371 • 21

Iker/NoticIA
Viewer • Updated Aug 6, 2024 • 850 • 654 • 1

Universidad del País Vasco Euskal Herriko Unibertsitatea
HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology

Adapt the lessons learned to the new chat-style LLM paradigm:

- Exploring the use of Machine Translation to generate instruction-tuning data for low-resource languages based on the already existing instruction-tuning datasets in high-resource languages.
- Synthetic data generation using LLMs: A model pre-trained with unstructured text from many languages and instruction-tuned in only a few high-resource languages may be able to generate synthetic data for all the languages it has been pre-trained on.
- Cultural adaptation of LLMs for low-resource languages.

Papers that are part of this thesis

- Iker García-Ferrero, Rodrigo Agerri, and German Rigau. Model and Data Transfer for Cross-Lingual Sequence Labelling in Zero-Resource Settings. (EMNLP 2022)
- Iker García-Ferrero, Rodrigo Agerri, and German Rigau. T-projection: High quality annotation projection for sequence labeling tasks. (EMNLP 2023)
- Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, Andrea Zaninello. Medical mT5: An Open-Source Multilingual Text-to-Text LLM for The Medical Domain. (LREC-COLING 2024)

Universidad del País Vasco Euskal Herriko Unibertsitatea
HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology

Closely Related Contributions

- Iker García-Ferrero, Rodrigo Agerri, and German Rigau. Benchmarking meta-embeddings: What works and what does not. (EMNLP 2021)
- Iker García-Ferrero, Jon Ander Campos, Oscar Sainz, Ander Salaberria, and Dan Roth. IXA/Cogcomp at SemEval-2023 Task 2: Context-enriched Multilingual Named Entity Recognition using Knowledge Bases. (SemEval 2023)
- Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, Eneko Agirre. GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction. (ICLR 2024)

Universidad del País Vasco Euskal Herriko Unibertsitatea

HiTZ Hizkuntza Teknologiako Zentroa Basque Center for Language Technology

Contributions that are not part of this thesis

- Salaberria, A., Campos, J. A., García-Ferrero, I., Fernandez de Landa, J. Itzulpen Automatikoko Sistemen Analisia: Genero Alborapenaren Kasua. (Ikergazte 2021)
- Fernandez de Landa, J., García-Ferrero, I., Salaberria, A., Campos, J. A. Twitterreko Euskal Komunitatearen Eduki Azterketa Pandemia Garaian. (Ikergazte 2021)
- García-Ferrero, I., Altuna, B., Álvez, J., Gonzalez-Dios, I., Rigau, G. This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models. (EMNLP 2023)
- Sainz, O., Campos, J. A., García-Ferrero, I., Etxaniz, J., Lopez de Lacalle, O., Agirre, E. NLP Evaluation in Trouble: On the Need to Measure LLM Data Contamination for each Benchmark. (EMNLP 2023)
- Fernandez de Landa, J., García-Ferrero, I., Salaberria, A., Campos, J. A. Uncovering Social Changes of the Basque Speaking Twitter Community During COVID-19 Pandemic. (SIGUL @ LREC-COLING 2024)
- García-Ferrero, I., Altuna, B. NoticIA: A Clickbait Article Summarization Dataset in Spanish. (PLN Journal 2024)
- Sainz, O., García-Ferrero, I., Jacovi, A., Campos, J. A., Elazar, Y., Agirre, E., Goldberg, Y., Chen, W.-L., Chim, J., Choshen, L., D'Amico-Wong, L., Dell, M., Fan, R.-Z., Golchin, S., Li, Y., Liu, P., Pahwa, B., Prabhu, A., Sharma, S., Silcock, E., Solonko, K., Stap, D., Surdeanu, M., Tseng, Y.-M., Udandarao, V., Wang, Z., Xu, R., Yang, J. Data Contamination Report from the 2024 CONDA Shared Task. (CONDA @ ACL 2024)

Cross-Lingual Transfer for Low-Resource Natural Language Processing

Transferencia crosslingüe para el Procesamiento del Lenguaje Natural

Iker García-Ferrero

Supervised by **German Rigau** and **Rodrigo Agerri**

HiTZ Zentroa - Ixa taldea
Euskal Herriko Unibertsitatea UPV/EHU

PhD Dissertation

February 12, 2025

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology