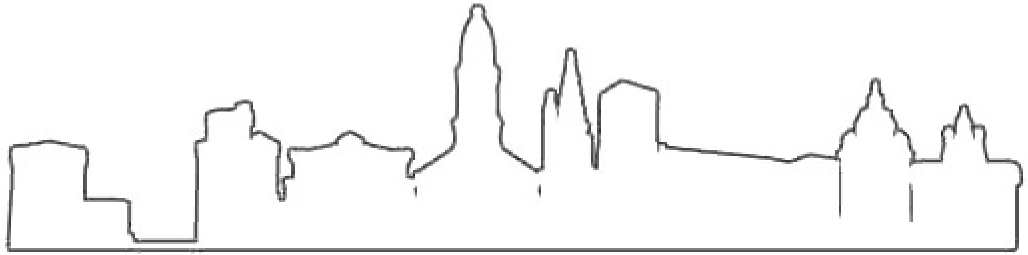
Sequence-to-Sequence Models for Extracting Information from Registration and Legal Documents

Ramon Pires Fábio C. de Souza Guilherme Rosa

Roberto A. Lotufo Rodrigo Nogueira

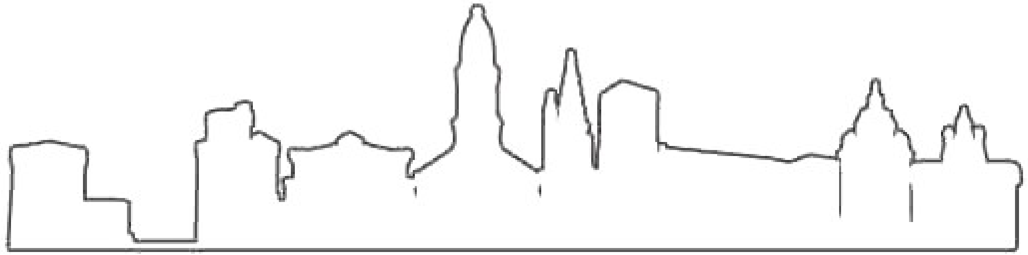
Agenda

1. Introduction
2. Methodology
3. Experimental Setup
4. Results
5. Conclusions

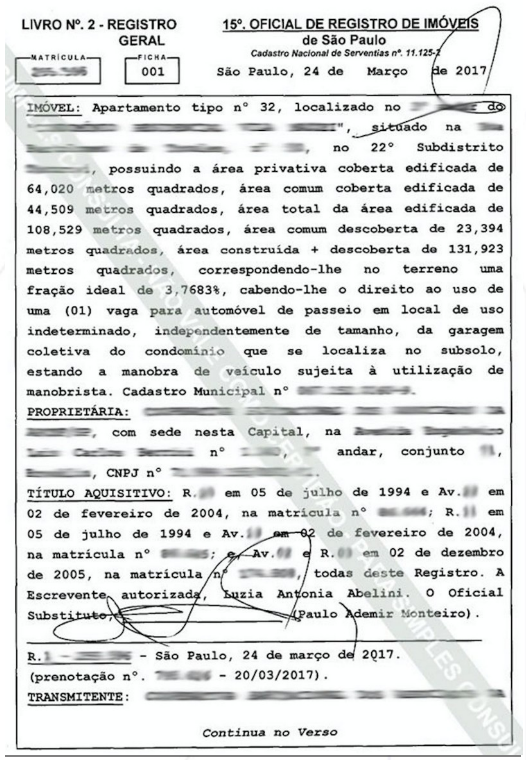


Agenda

1. Introduction
2. Methodology
3. Experimental Setup
4. Results
5. Conclusions



Introduction



LIVRO Nº. 2 - REGISTRO GERAL

15º. OFICIAL DE REGISTRO DE IMÓVEIS
de São Paulo
Cadastro Nacional de Serventias nº. 11.125-2

MATRÍCULA [redacted] FICHA 001

São Paulo, 24 de Março de 2017

IMÓVEL: Apartamento tipo nº 32, localizado no [redacted] do [redacted]", situado na [redacted], no 22º Subdistrito [redacted], possuindo a área privativa coberta edificada de 64,020 metros quadrados, área comum coberta edificada de 44,509 metros quadrados, área total da área edificada de 108,529 metros quadrados, área comum descoberta de 23,394 metros quadrados, área construída + descoberta de 131,923 metros quadrados, correspondendo-lhe no terreno uma fração ideal de 3,7683%, cabendo-lhe o direito ao uso de uma (01) vaga para automóvel de passeio em local de uso indeterminado, independentemente de tamanho, da garagem coletiva do condomínio que se localiza no subsolo, estando a manobra de veículo sujeita à utilização de manobrista. Cadastro Municipal nº [redacted].

PROPRIETÁRIA: [redacted], com sede nesta Capital, na [redacted] nº [redacted], [redacted] andar, conjunto [redacted], [redacted], CNPJ nº [redacted].

TÍTULO AQUISITIVO: R.[redacted] em 05 de julho de 1994 e Av.[redacted] em 02 de fevereiro de 2004, na matrícula nº [redacted]; R.[redacted] em 05 de julho de 1994 e Av.[redacted] em 02 de fevereiro de 2004, na matrícula nº [redacted]; e Av.[redacted] e R.[redacted] em 02 de dezembro de 2005, na matrícula nº [redacted], todas deste Registro. A Escrevente autorizada, Luzia Antonia Abelini. O Oficial Substituto, (Paulo Ademir Monteiro).

R.[redacted] - São Paulo, 24 de março de 2017.
(prenotação nº. [redacted] - 20/03/2017).
TRANSMITENTE: [redacted]

Continua no Verso

Q: What is the state?

A: SP

Q: What is the county?

A: São Paulo

Q: What is the office?

A: 15º Oficial de Registro de Imóveis

Q: What is the private area?

A: 64,020 m2

Banco do Nordeste

Proposta de Cadastro
Pessoa Física

Banco do Nordeste para fazer a diferença na sua vida.

Para que o conheçamos melhor, faz-se necessário o preenchimento completo deste formulário, rubricando-o em todas as suas páginas e assinando-o ao final, em campo destinado a este fim.

Seja bem-vindo ao Banco que faz a diferença na vida de todos os nordestinos.

1Agência Responsável
[redacted]

I - IDENTIFICAÇÃO

Nome: [redacted]
Como gostaria de ser chamado?: [redacted]

CPF: [redacted]
Nº Documento de Identificação: [redacted]
Órgão Emissor: SSP
UF: CE
Data da Emissão: [redacted]

Data de Nascimento: [redacted]
Sexo: ☒ M ☐ F
País de Nascimento: BRASIL
Possui Múltipla Cidadania? ☐ Sim ☒ Não

Nacionalidade(s): BRASIL
Naturalidade (se país de nascimento for Brasil): [redacted]
UF Nascimento (se país de nascimento for Brasil): CE
Possui Green Card? ☐ Sim ☒ Não

Possui residência fiscal em outro país que não o Brasil? ☐ Sim (Informar o País) ☒ Não
Caso o país de residência fiscal não seja o Brasil, informar o NIF (Número de Identificação Fiscal)
Naturalizado? ☐ Sim ☒ Não

Nome do Pai: [redacted]
Nome da Mãe: [redacted]

Estado Civil: ☒ Solteiro ☐ Casado ☐ Separado Judicialmente ☐ Divorciado ☐ Viúvo
Regime de Bens: ☐ Comunhão Parcial ☐ Comunhão Universal ☐ Participação final nos aquestos ☐ Separação Total

Possui União Estável? ☐ Sim ☒ Não
Nome do Cônjuge ou Companheiro(a)

CPF do Cônjuge ou Companheiro(a)
Renda Bruta Mensal do Cônjuge ou Companheiro (a)

Grau de Instrução: ☐ Analfabeto ☐ Ensino Fundamental Incompleto ☐ Ensino Fundamental Completo ☐ Ensino Médio Incompleto ☐ Ensino Médio Completo ☐ Superior Incompleto ☒ Superior Completo ☐ Pós-graduação ☐ Mestrado ☐ Doutorado
Cor ou Raça (categorias IBGE): ☐ Indígena ☐ Preta ☐ Amarela ☐ Parda ☒ Branca

Endereço Residencial (Logradouro, nº, complemento, bairro, cidade, UF, CEP): [redacted]

Telefone Residencial: [redacted]
Celular: [redacted]
Fax: ( )
E-mail: [redacted]
Rede Social

Endereço Comercial (Logradouro, nº, complemento, bairro, cidade, UF, CEP): [redacted]

Q: What is the full name?

A: *(anonymized)*

Q: What is the address?

A: *(anonymized)*

Q: What is the zip code?

A: *(anonymized)*

Q: What is the identity number?

A: *(anonymized)*

Introduction

Current commercial information extraction (IE) systems consist of individual modules controlled by manually defined rules

In production pipelines, the requirements and specifications often change

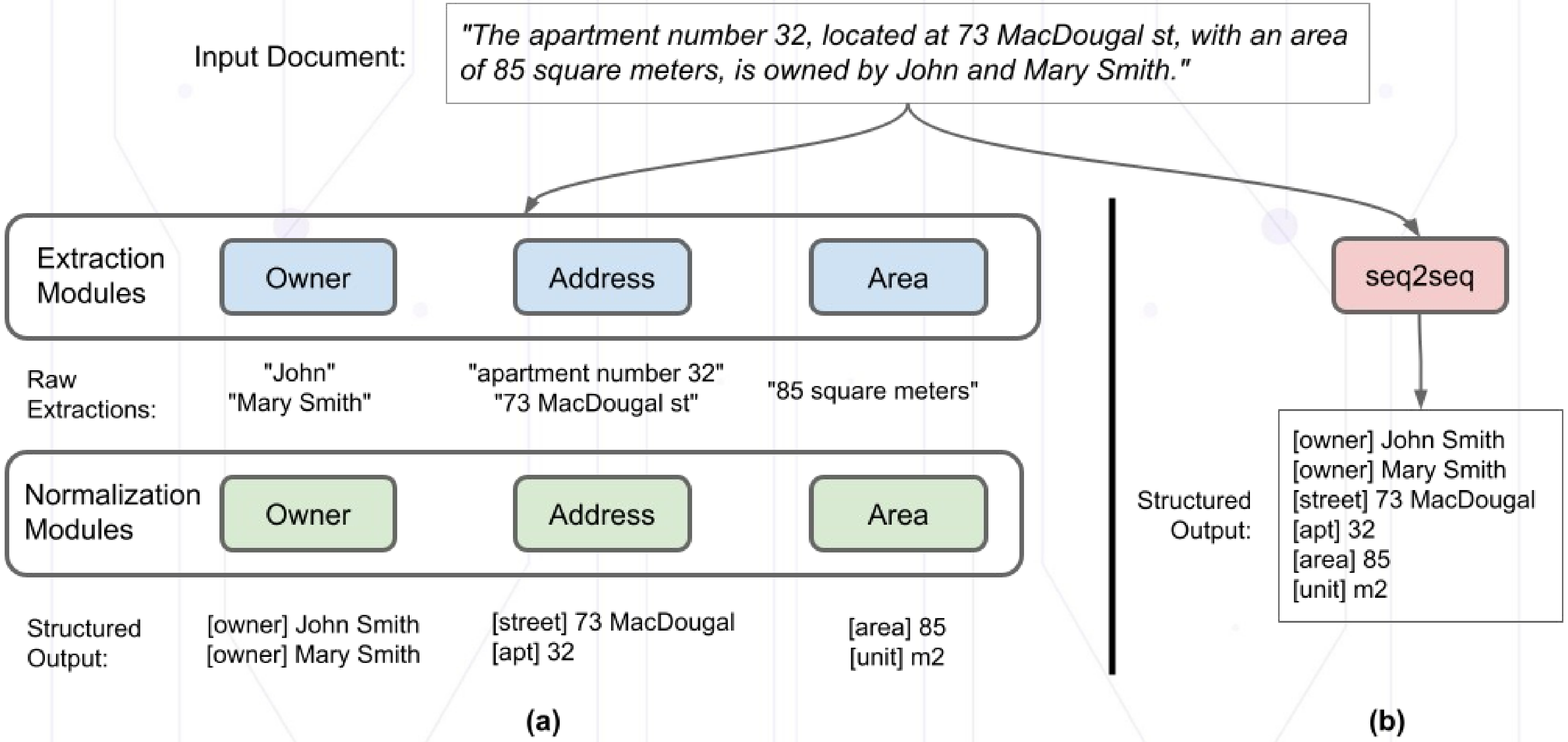
This leads to higher maintenance costs due to an ever larger number of individual components

Our Proposal

We study the viability of a framework for information extraction based on a **single** sequence-to-sequence model for **extracting** and **processing** information from **legal and registration documents**

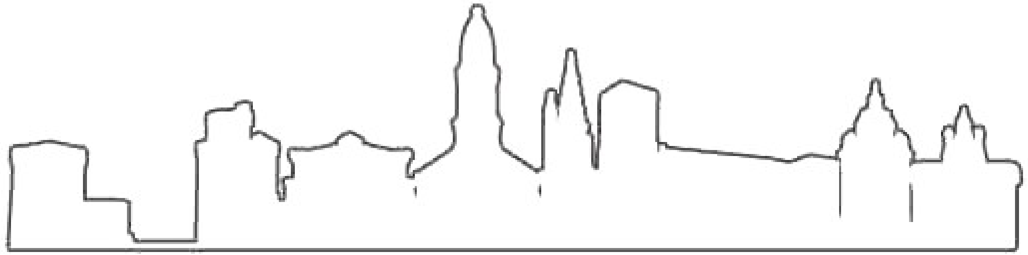
- ➔ a single model needs to be trained and maintained
- ➔ can be shared by multiple projects with different requirements and types of documents

Our Proposal



Agenda

1. Introduction
2. Methodology
3. Experimental Setup
4. Results
5. Conclusions



Questions and Answers

Our method uses **questions** coupled with contexts as input and **answers** as output

For consistency, we strive to formulate questions as general as possible

We also format the answers by adding **clues** for each category of information

➔ important to structure the response

➔ useful for designing compound answers

Q&A: Simple QAs

context: Apartment type nº 32, located on the 10th floor of the Central Building, situated at 1208 Santos Dumont St., having a private covered built area of 64,020 square meters, a common covered built area of 44,509 square meters, a total built area of 108,529 square meters...

Q: `What is value of the private area?`

A: **`[value]:`** `64.02`

Q: `What is the unit of measure of the private area?`

A: **`[unit]:`** `square meters`

IAPR

neuralmind.ai

Q&A: Compound QAs

Some subsets of fields are often closely related or even appear connected

The classical pipeline is cumbersome as it requires the model to analyze each document oftentimes for extracting information of the same scope

By using **compound QAs**, a set of questions for extracting individual information are replaced by a single question

Q&A: Compound QAs

context: Apartment type nº 32, located on the 10th floor of the Central Building, situated at 1208 Santos Dumont St., having a private covered built area of 64,020 square meters, a common covered built area of 44,509 square meters, a total built area of 108,529 square meters...

Q: `What is the private area?`

A: `[value]: 64.02 [unit]: square meters`

Q&A: Sentence IDs

One way to monitor the quality of predictions is to know the **location** in the input text from which the information was extracted

However, the location cannot be trivially inferred from the output of seq2seq models

To address this limitation, we propose the use of **sentinel tokens** that allow the generative model to reveal the location of its prediction in the original sequence

Q&A: Sentence IDs

context: Apartment type nº 32, located on the 10th floor of the Central Building, situated at 1208 Santos Dumont St., having a private covered built area of 64,020 square meters, a common covered built area of 44,509 square meters, a total built area of 108,529 square meters...

Q: `What is the private area?`

A: **[value]:** `64.02` **[unit]:** `square meters`

Q&A: Sentence IDs

context: [SENT1] Apartment type nº 32, [SENT2] located on the 10th floor of the Central Building, [SENT3] situated at 1208 Santos Dumont St., [SENT4] having a private covered built area of 64,020 square meters, [SENT5] a common covered built area of 44,509 square meters, [SENT6] a total built area of 108,529 square meters...

Q: `What is the private area?`

A: **[value]:** `64.02` **[unit]:** `square meters`

Q&A: Sentence IDs

context: [SENT1] Apartment type nº 32, [SENT2] located on the 10th floor of the Central Building, [SENT3] situated at 1208 Santos Dumont St., [SENT4] having a private covered built area of 64,020 square meters, [SENT5] a common covered built area of 44,509 square meters, [SENT6] a total built area of 108,529 square meters...

Q: `What is the private area?`

A: `[SENT4] [value]: 64.02 [SENT4] [unit]: square meters`

Q&A: Sentence IDs

context: [SENT1] Apartment type nº 32, [SENT2] located on the 10th floor of the Central Building, [SENT3] situated at 1208 Santos Dumont St **[SENT4] having a private covered built area of *64,020 square meters*,** [SENT5] a common covered built area of 44,509 square meters, [SENT6] a total built area of 108,529 square meters...

Q: `What is the private area?`

A: `[SENT4] [value]: 64.02 [SENT4] [unit]: square meters`

Q&A: Canonical format

Often, certain types of information appear in a document in a variety of formats

- ➔ **23 May 2022**
- ➔ **May 23, 2022**
- ➔ **23/05/2022**
- ➔ **2022/05/23**
- ➔ **23-05-2022**
- ➔ **2022-05-23**

Our IE system is able to directly extract those particular fields in canonical format

However, with canonical formats, the use of **sentence IDs may not be enough** for locating the extracted information in the original document

Q&A: Canonical format

context: [SENT1] Apartment type nº 32, [SENT2] located on the 10th floor of the Central Building, [SENT3] situated at 1208 Santos Dumont St., [SENT4] having a private covered built area of 64,020 square meters, [SENT5] a common covered built area of 44,509 square meters, [SENT6] a total built area of 108,529 square meters...

Q: `What is the private area?`

A: `[SENT4] [value]: 64.02 [SENT4] [unit]:` m^2

Q&A: Canonical format

context: [SENT1] Apartment type nº 32, [SENT2] located on the 10th floor of the Central Building, [SENT3] situated at 1208 Santos Dumont St., [SENT4] having a private covered built area of 64,020 square meters, [SENT5] a common covered built area of 44,509 square meters, [SENT6] a total built area of 108,529 square meters...

Q: `What is the private area and how does it appear in text?`

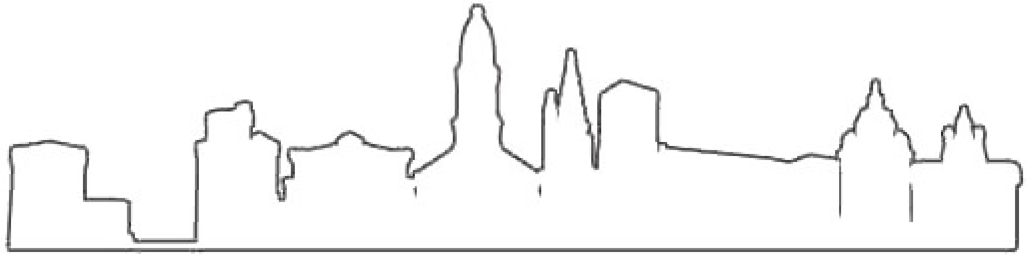
A: `[SENT4] [value]: 64.02 [SENT4] [unit]: m² [text] square meters`

IAPR

neuralmind.ai

Agenda

1. Introduction
2. Methodology
3. **Experimental Setup**
4. Results
5. Conclusions



Models

PTT5:

T5-base model pretrained on a large Brazilian Portuguese corpus

PTT5-Legal:

PTT5 pretrained on legal documents in Portuguese

PTT5-QA:

PTT5 pre-finetune on the SQuAD v1.1 dataset

PTT5-Legal-QA:

In-domain pretraining followed by task-aware pre-finetuning

Datasets

NM-Property:

Property registres (legal domain)

NM-Certificates:

Certificates (legal domain)

NM-Legal:

Legal notices (legal domain)

NM-Forms:

Forms (registration domain)

Datasets

Table 1: Statistics of the datasets.

Dataset	Chars/doc	Fields	Train	Valid	Test
NM-Property	3011.53	17	3191	799	242
NM-Certificates	4914.39	10	760	191	311
NM-Publications	1895.76	3	1600	401	500
NM-Forms	1917.14	25	240	60	282
Total	–	55	5791	1451	1334

Metrics

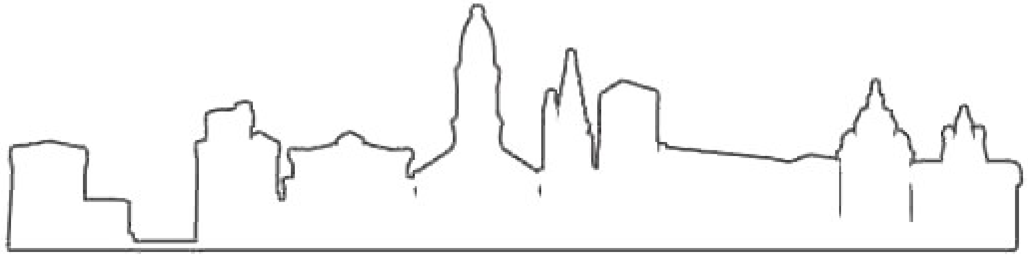
Exact matching (EM): accuracy

F1-measure (F1): token-based harmonic mean of precision and recall

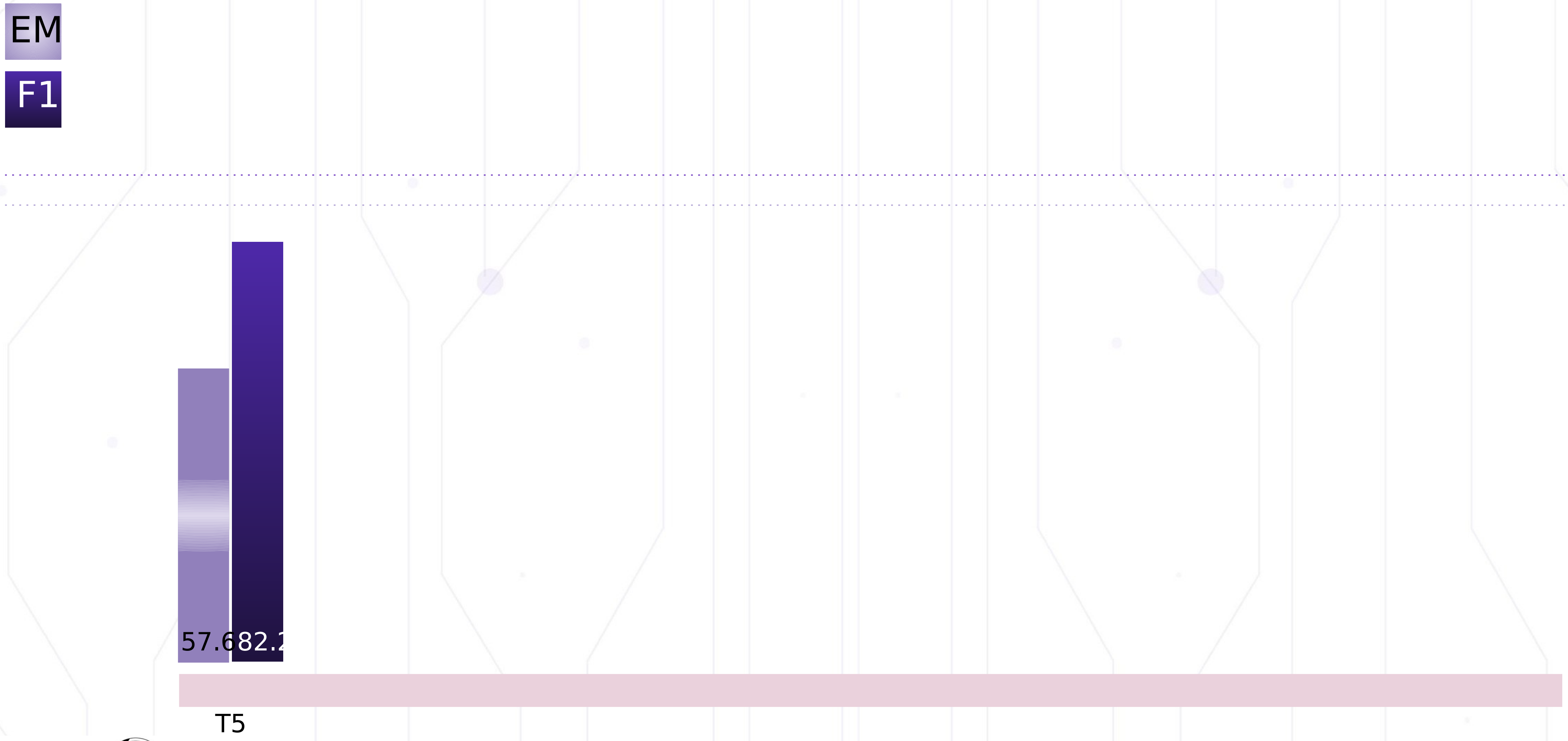
Before computing the metrics, both label and prediction sentences are normalized by converting to **lower case**, removing **double spaces** and **punctuation**

Agenda

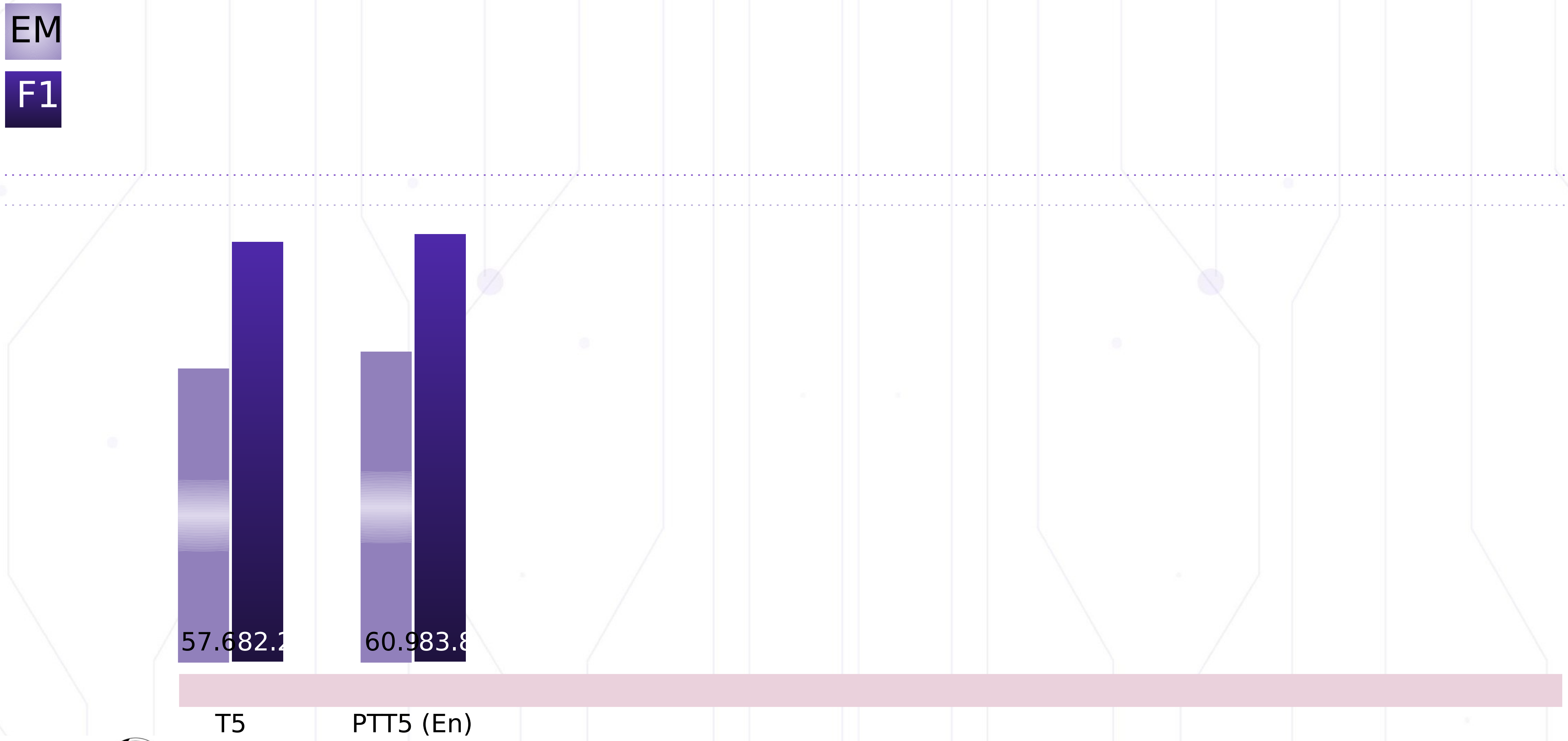
1. Introduction
2. Methodology
3. Experimental Setup
4. **Results**
5. Conclusions



Results



Results

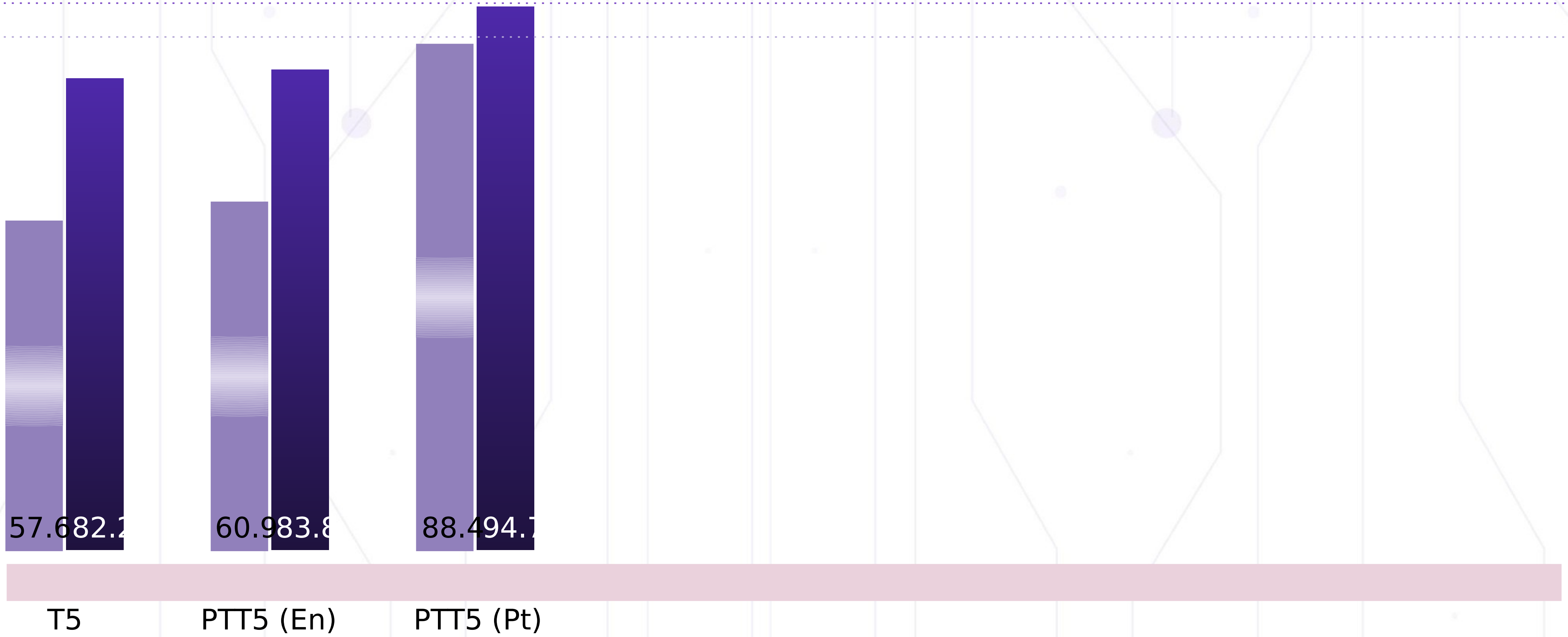


Results

EM

F1

The adoption of Portuguese tokenizer provided an error reduction in EM of **70.3%** over the previous experiment that used the same pretraining dataset

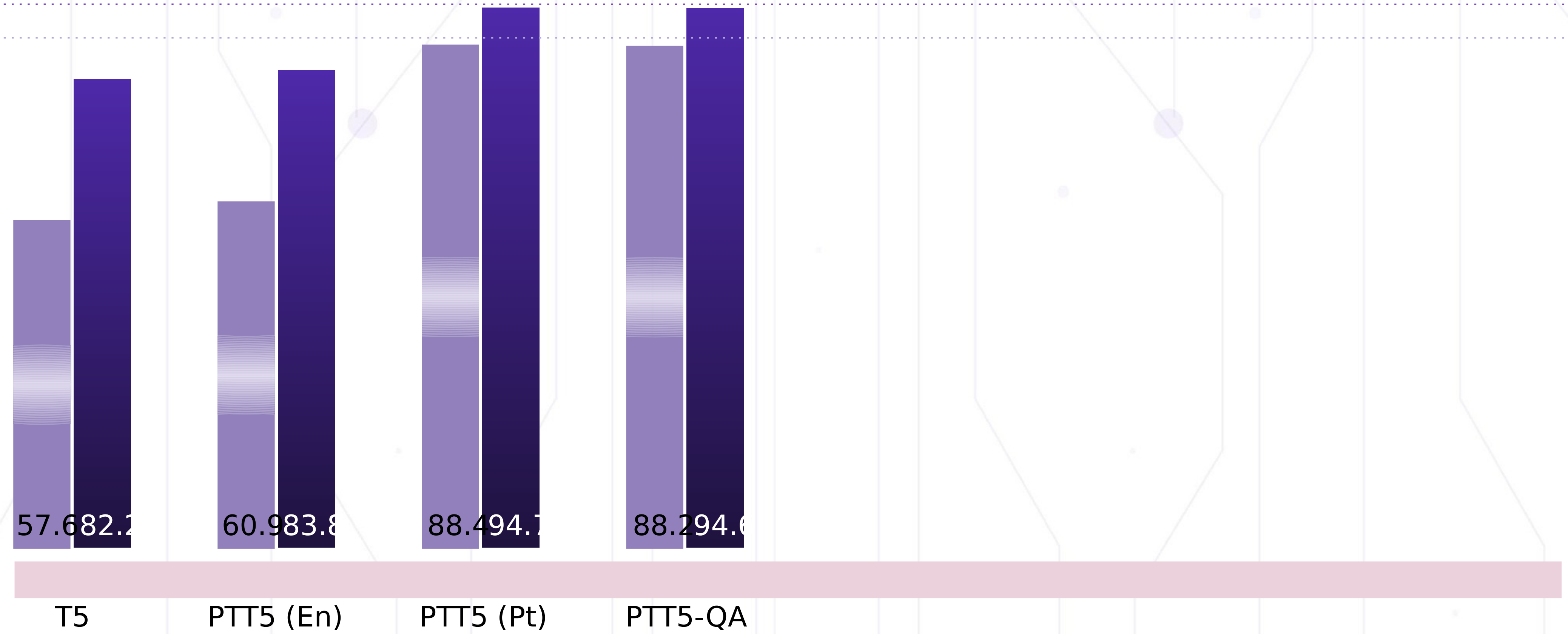


Results

EM

F1

Adapting the model for QA on the SQuAD dataset **did not provide** improvements over the large-scale pretraining

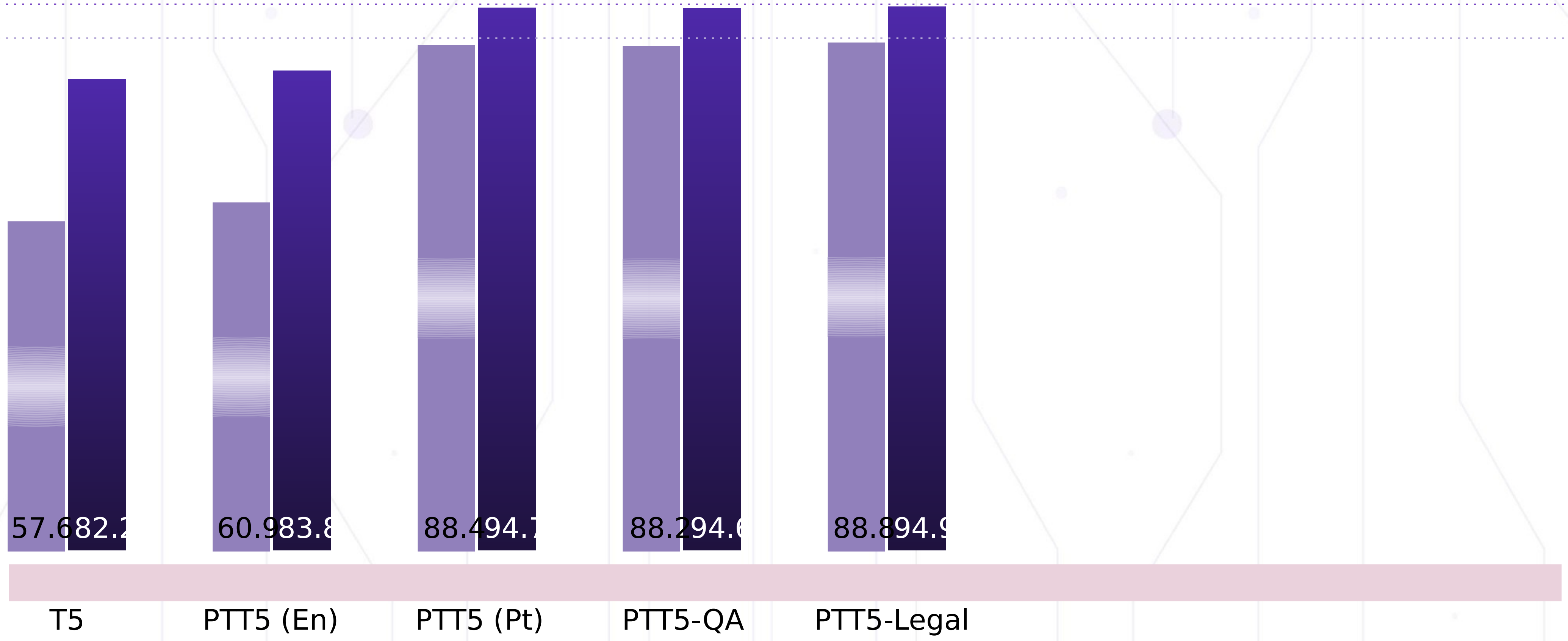


Results

EM

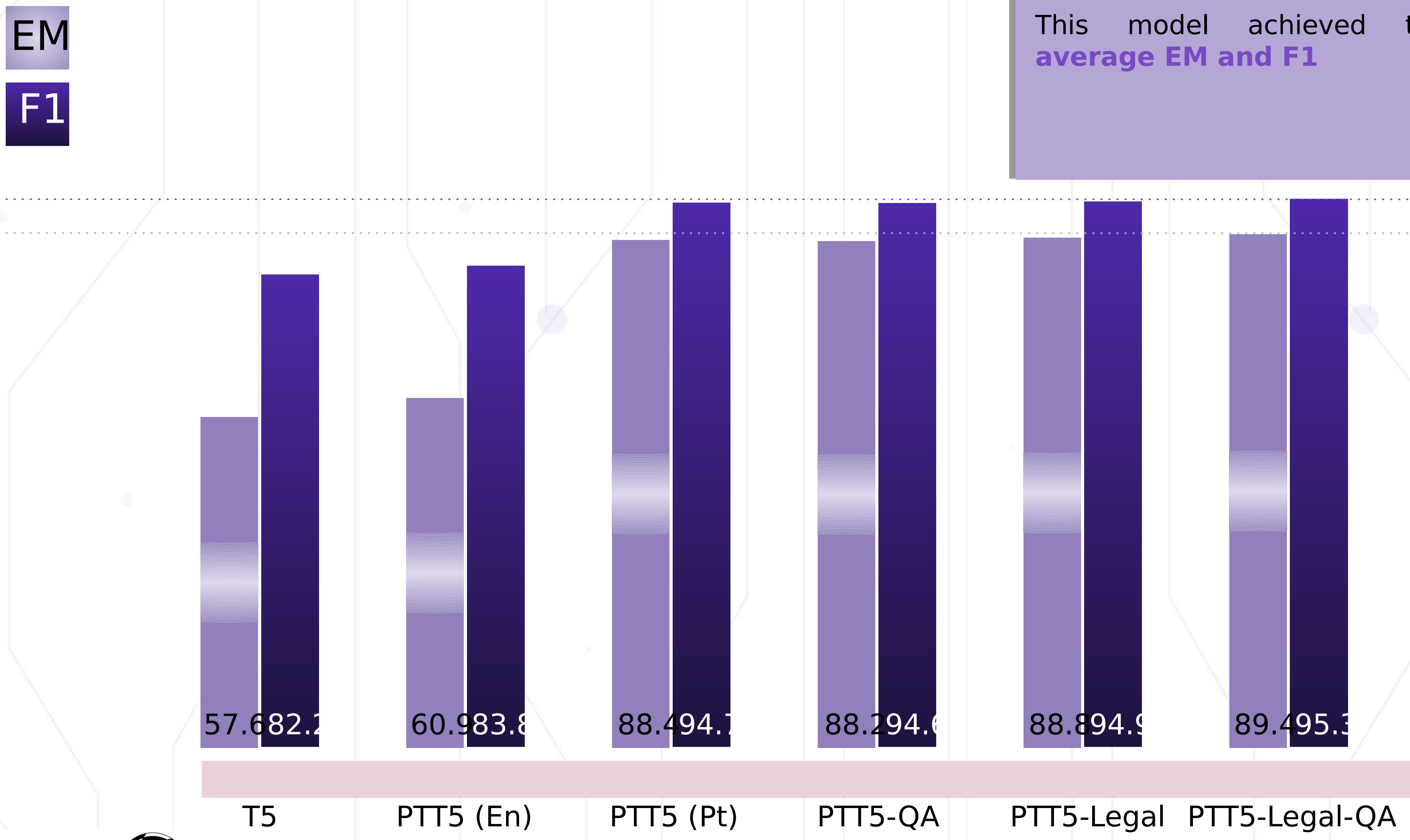
F1

The unsupervised pretraining brought the best result over the **NM-Properties** dataset, and a minor improvement on the average of the four datasets



Results

This model achieved the **best average EM and F1**



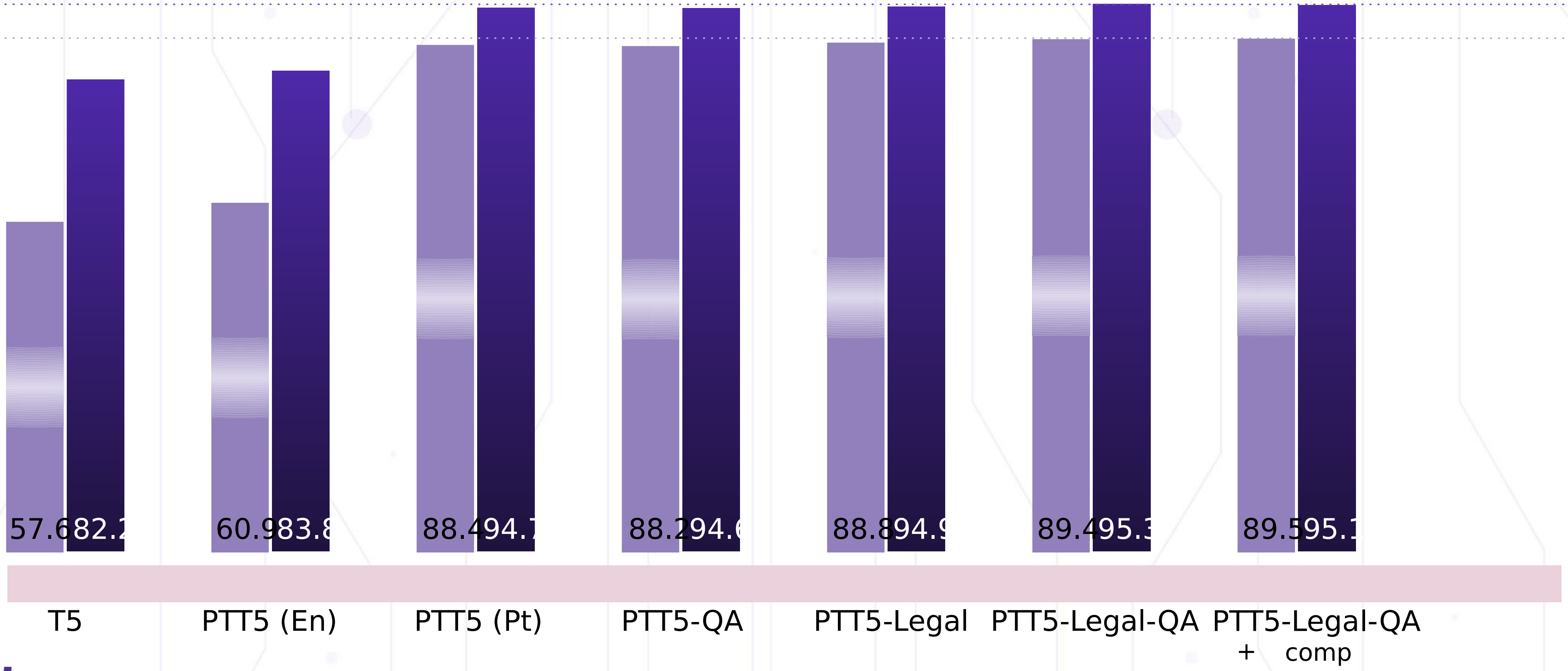
Results

EM

F1

Although **superior** to results using individual QAs, the average results are comparable

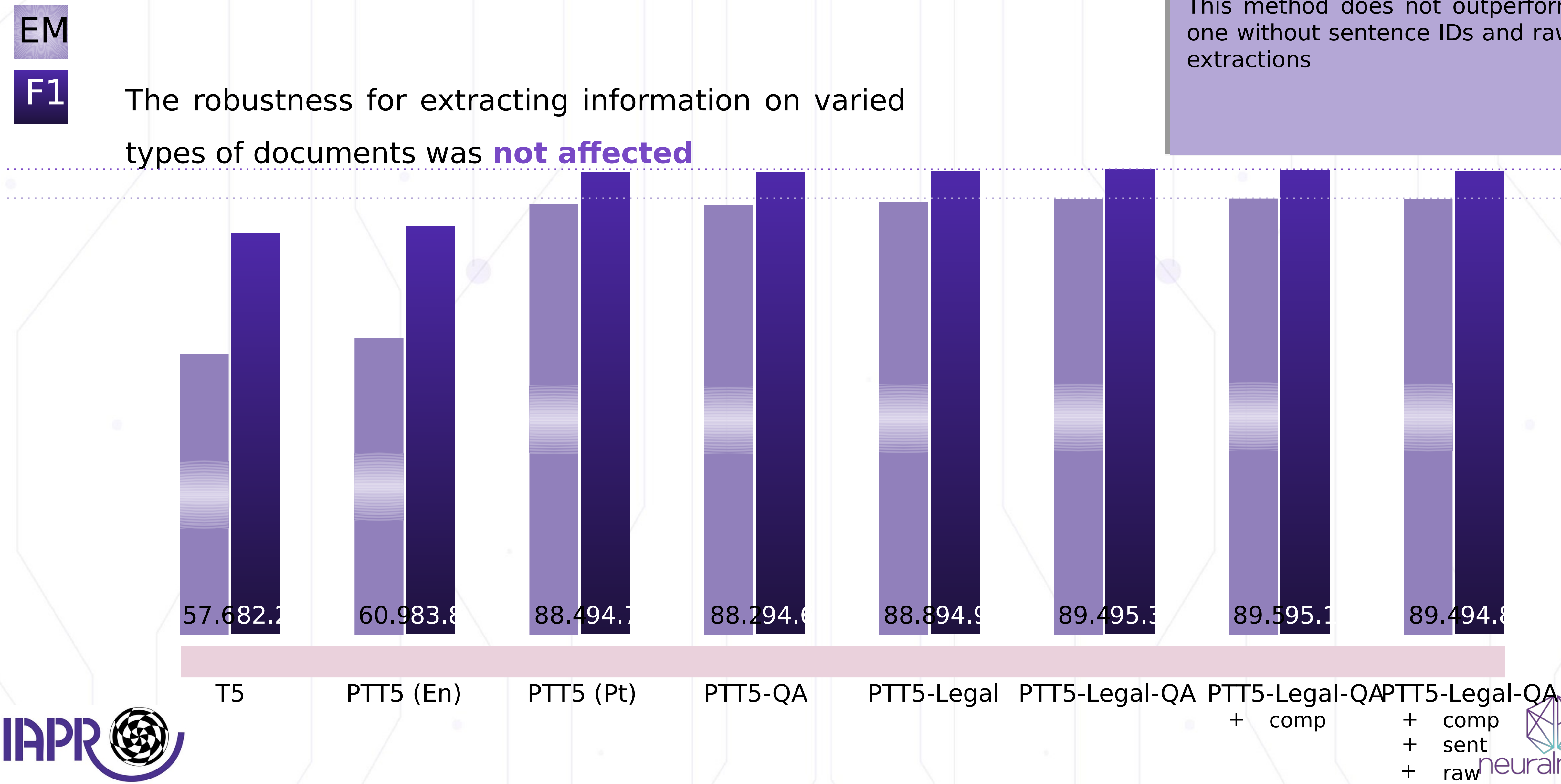
Compound QAs yields **better results** than using individual QAs, reaching superior performance for three datasets



Results

This method does not outperform the one without sentence IDs and raw-text extractions

The robustness for extracting information on varied types of documents was **not affected**



Comparison with BERT

We compare seq2seq models with the classical **named entity recognition**

The dataset was filtered out, retaining **38%** of the documents and **42 of the 55** fields.

Table 3: NER ablation results .

Model	Params	Precision	Recall	F1-score (micro)
BERT-Large	330M	90.2	92.9	91.5
T5-base (ours)	220M	91.6	89.6	90.6

Comparison with BERT

We compare seq2seq models with the classical **named entity recognition**

The dataset was filtered out, retaining **38%** of the documents and **42 of the 55** fields.

Table 3: NER ablation results .

Model	Params	Precision	Recall	F1-score (micro)
BERT-Large	330M	90.2	92.9	91.5
T5-base (ours)	220M	91.6	89.6	90.6

The removed cases would be **impossible** for NER **(false negatives)**

Comparison with BERT

We compare seq2seq models with the classical **named entity recognition**

The dataset was filtered out, retaining **38%** of the documents and **42 of the 55** fields.

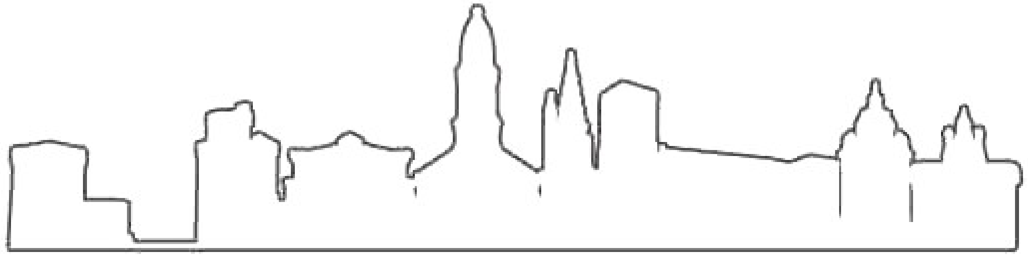
Table 3: NER ablation results .

Model	Params	Precision	Recall	F1-score (micro)
BERT-Large	330M	90.2	92.9	91.5
T5-base (ours)	220M	91.6	89.6	90.6

The removed cases would be **impossible** for NER **(false negatives)**

Agenda

1. Introduction
2. Methodology
3. Experimental Setup
4. Results
5. Conclusions



Conclusions

We validated the use of a single seq2seq model for **extracting information** from four different classes of legal and registration documents in Portuguese

The model is trained end-to-end to output **structured** text, thus replacing parts of rule-based normalization and post-processing steps of a classical pipeline

Models

Language (Portuguese) pretraining and tokenization are **the most important adaptations to increase effectiveness**

Pretraining on in-domain (legal) text and pre-finetuning on a large question-answering dataset **marginally improve results**

Sentence IDs and Canonical

Sentence IDs

We propose a method to **align answers with the input text**, thus allowing seq2seq models to be more easily monitored and audited in IE pipelines

Canonical Format

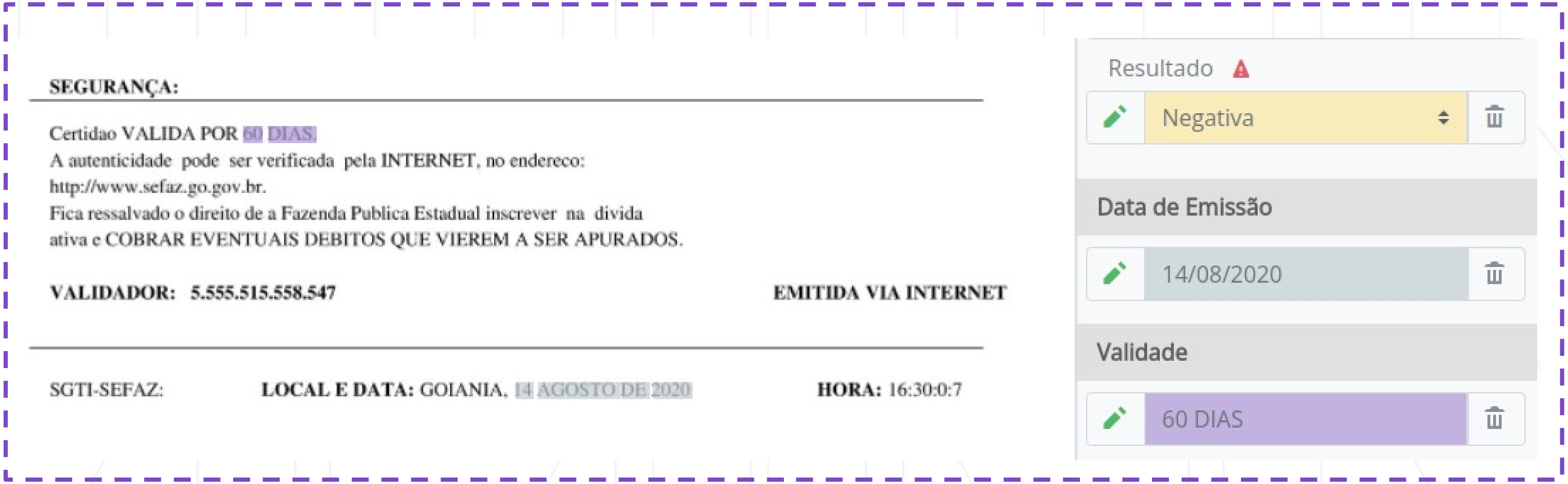
The model is **capable of extracting canonical formats** for dates that can originally appear in the document in various formats

```
[SENT14] [Issue Date]: 14/01/2020

[text] 14 days of the month of August of the year 2020
```

example in which the model fails to convert raw text into canonical text (translated from portuguese)

Sentence IDs and Canonical



SEGURANÇA:

Certidao VALIDA POR 60 DIAS.
A autenticidade pode ser verificada pela INTERNET, no endereco:
http://www.sefaz.go.gov.br.
Fica ressalvado o direito de a Fazenda Publica Estadual inscrever na divida ativa e COBRAR EVENTUAIS DEBITOS QUE VIEREM A SER APURADOS.

VALIDADOR: 5.555.515.558.547 **EMITIDA VIA INTERNET**

SGTI-SEFAZ: **LOCAL E DATA:** GOIANIA, 14 AGOSTO DE 2020 **HORA:** 16:30:0:7

Questions?

Source code

https://github.com/neuralmind-ai/information-extraction-t5