Generating Text from RDF Data

Oral and Written Language Processing
(GCED-UPC) Spring Semester

David Bergés, Roser Cantenys, Roger Creus, Oriol Domingo
Professor: José Adrián Fonollosa

1) Problem

2) System Architecture

3) Experiments

4) Results

5) Next Steps

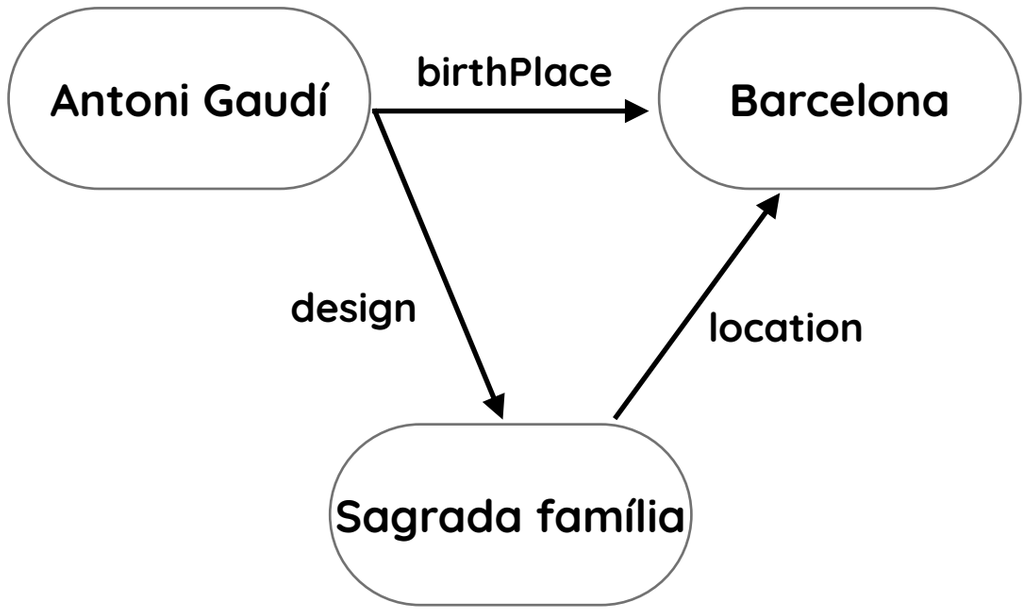
Problem - Contextualization

- **Knowledge Base (KB):** large source of information represented in a structured way.

- **Resources Description Frameworks (RDF):** Base of the information structure which consists in 3 elements and establishes relations between them.

    - Predicate (relation), subject and object (entities)

Problem - Example



(a) Knwoledge graph.

RDF Triples	<Antoni Gaudí, design, Sagrada Família> <Antoni Gaudí, birthPlace, Barcelona> <Sagrada Família, location, Barcelona>

(b) Knwoledge base and its RDF triples.

Natural Sentence	Antoni Gaudí was born in Barcelona, the city where he designed Sagrada Família.

(c) Corresponding natural language description.

Problem - Description

- **Two main tasks**
    - RDF to Text generation
    - Text to RDF semantic parsing

Problem - Formal Definition

- **Input:** KB denoted as a set of RDFs, $\mathscr{K} := \{r_1, \ldots, r_n\}$ and each RDF is defined as $\langle s_i, p_i, o_i \rangle$.
- **Output:** sequence of sentences $\mathcal{S}$, each sentence is a sequence of words $[w_1, \ldots, w_m]$ which should be grammatically correct and should also contain all the information present in $\mathscr{K}$.

Problem - Objectives

- **End-to-end Architectures**
    - Encoder-Decoder Convolutional
    - Encoder-Decoder Transformer
- **Model Enhancement**
    - Byte-Pair Encoding
    - Back Translation
    - Embeddings

System Architecture - Approach

- **Natural Language Generation (NLG)**

  · Data to Text - RDF to Text

  · Text to Text - MachineTranslation

Albert Gatt and Emiel Krahmer:
Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation

System Architecture - Preprocessing

- **Delexicalisation:** Entity name to Entity type

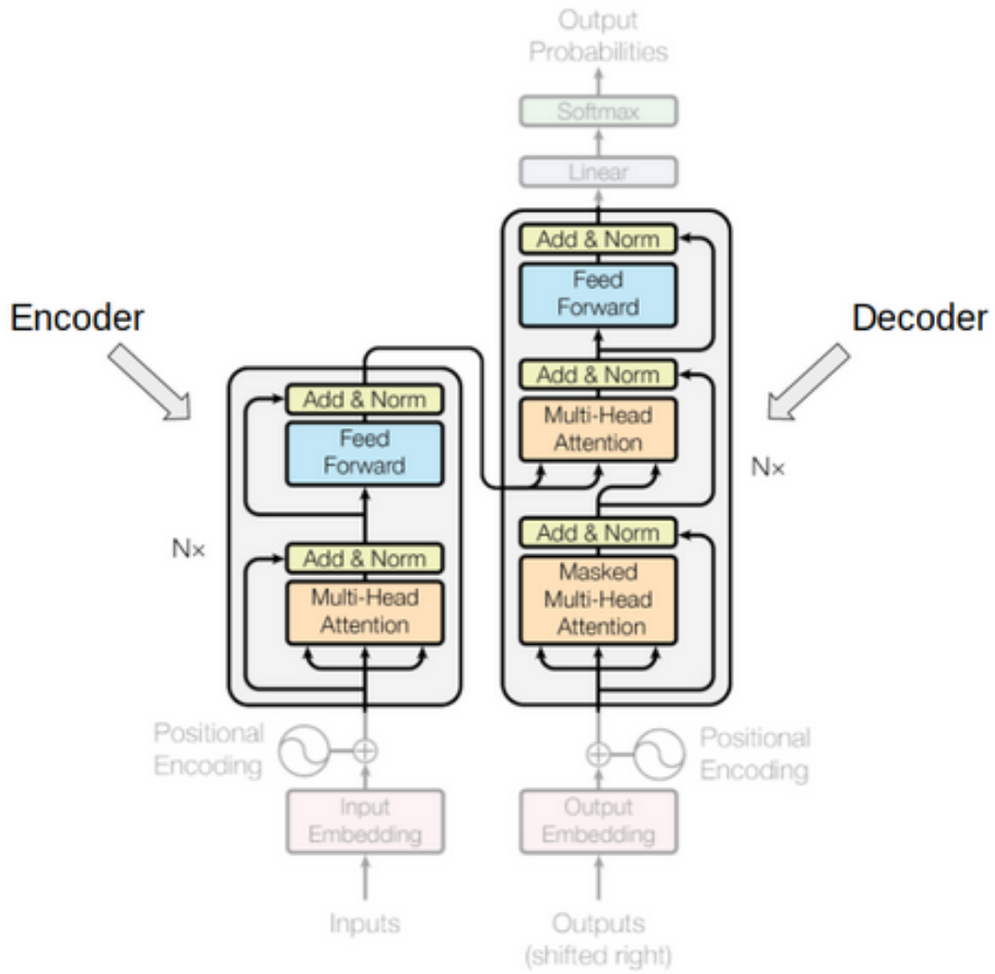
  (Rome, capital of, Italy) —> (CITY, CAPITALOF, COUNTRY)

- **Moses Tokenization**

- **Byte-Pair Encoding**

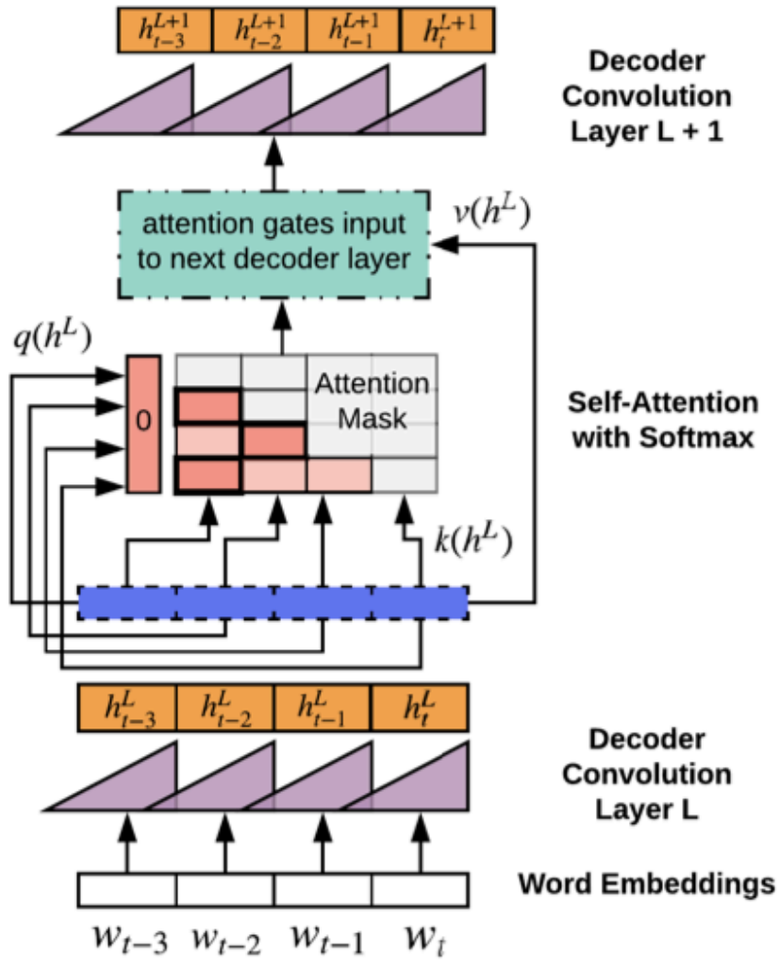
System Architecture - Models

Transformer



Ashish Vaswani et al. :
Attention Is All You Need

Convolutional



Angela Fan et al. :
Hierarchical Neural Story Generation

System Architecture - Postprocessing

- **Relexicalisation:** Entity type to Entity name

  CITY is the CAPITALOF COUNTRY —> Rome is the capital of Italy

- **Moses Detokenization**

- **Byte-Pair Decoding**

Experiments - Byte-Pair Encoding & Embeddings

- **Byte-Pair Encoding**
- **BPE + Learned Embeddings vs Pretrained Embeddings**

Experiments - Back Translation

Back Translation Model

Transformer Model

Text —> RDF

Predict using Monolingual

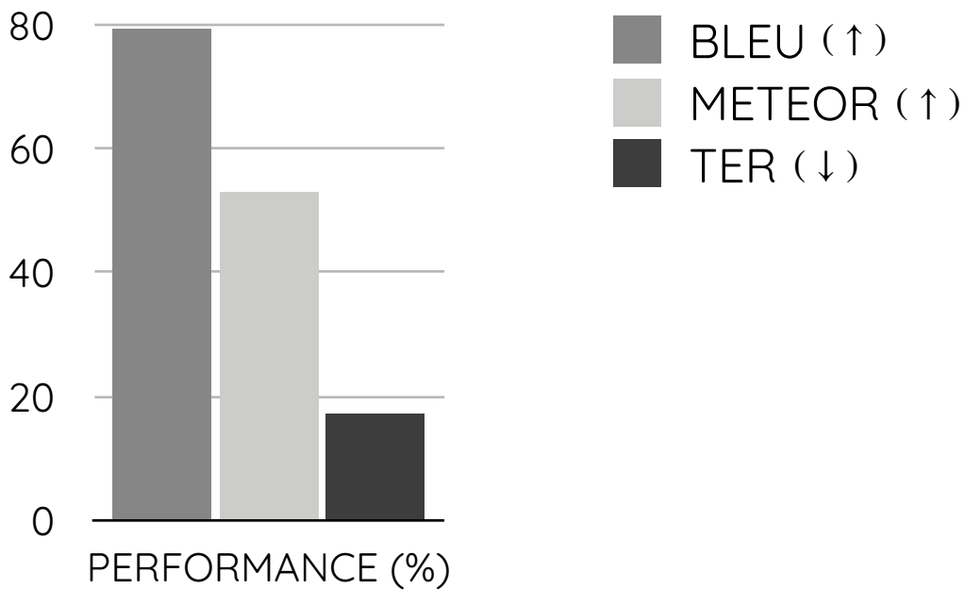
Wikipedia Data Base

New Text-RDF

Experiments - Back Translation

Back Translation Model

BPE + Lexicalise



Predict using Monolingual

Preprocessing DB

Poor synthetic data

↓

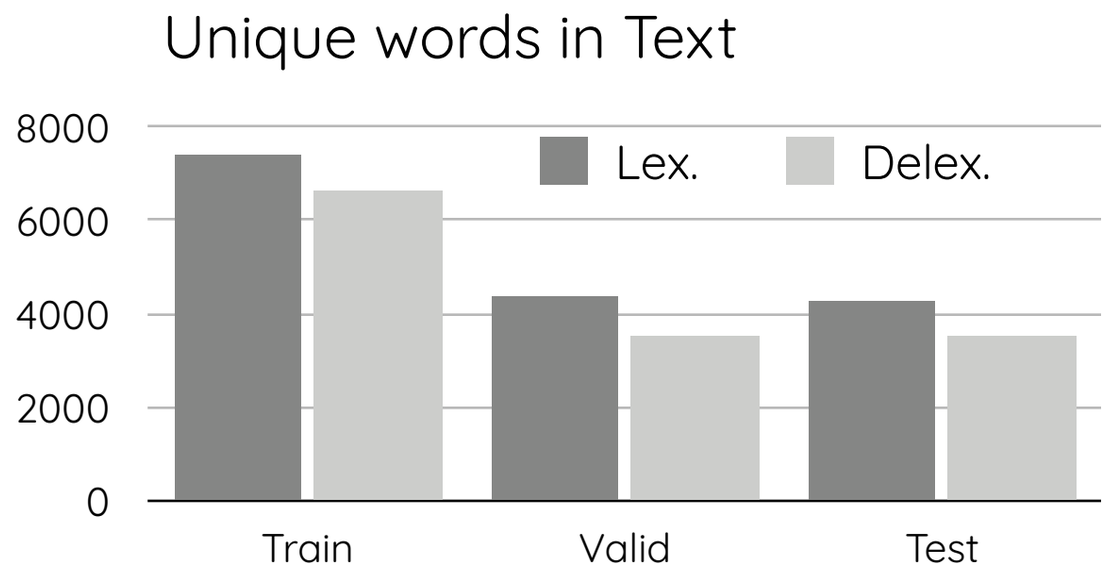
Remove BPE

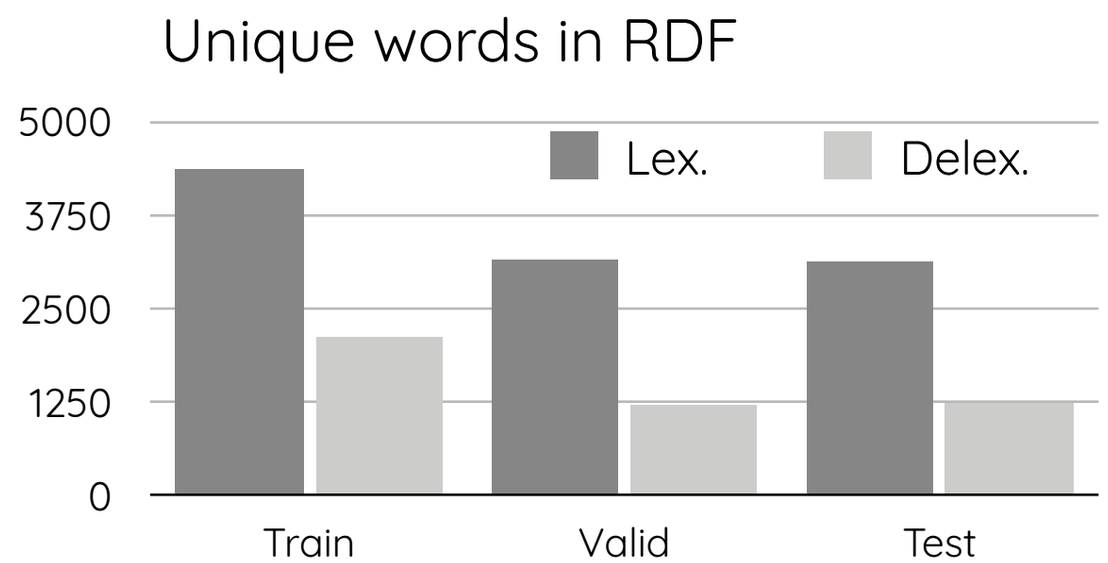
Filter Monolingual text

Pretrained embeddings

Results - Data

	Instances
Train	34.338
Valid	4.313
Test	4.222

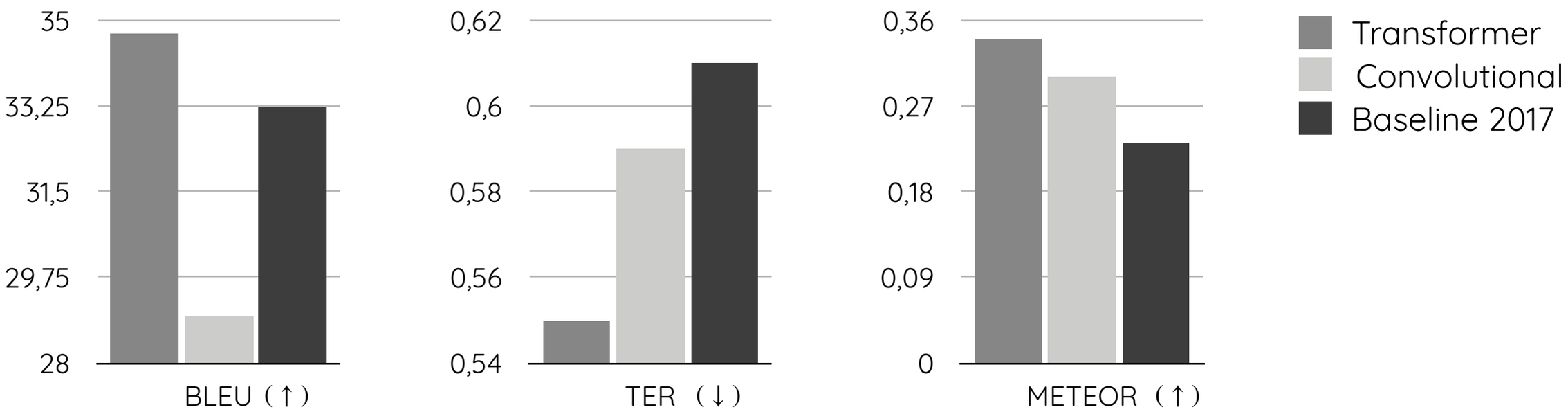




Results - Hyperparameter Tuning

- Grid Search on Transformer > 100 models
- Small Search on Convolutional 3 models

 # PARAMETERSS T: aprox.6M C: aprox.30M

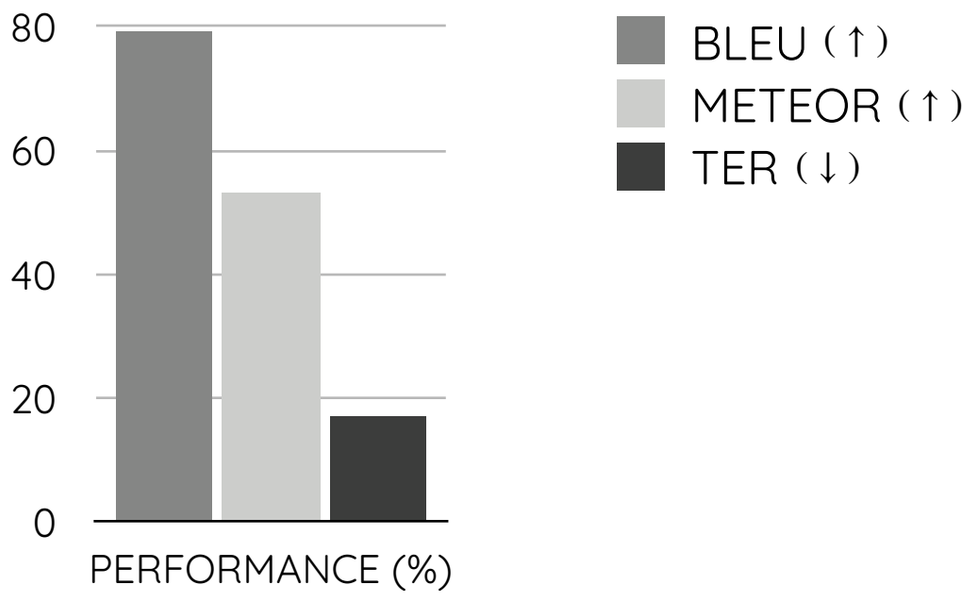


Results - System Prediction

Input	Output
Arròs negre country **Spain**	**Arròs negre** is a food found in **Catalonia** , **Spain** . The main ingredients of **Arròs negre** are White rice , cuttlefish or squid , cephalopod ink , cubanelle peppers
Arròs negre mainIngredients White rice , cuttlefish or squid , cephalopod ink , cubanelle peppers	
Arròs negre region **Catalonia**	
Arròs negre ingredient **Cuttlefish**	

Results - Back Translation

Back Translation Performance



Generated Data

MOVE THIS SLIDE TO EXPERIMENTS

Next Steps - Short Term

- **Training Phase**
  - More solid results
  - Use BT generated Data
- **Experiments**
  - Influence of BPE
  - Comparative results of embeddings

Next Steps - Long Term

- **Submit Results** WebNLG Challenge 2020

Generating Text from RDF Data

Oral and Written Language Processing
(GCED-UPC) Spring Semester

David Bergés, Roser Cantenys, Roger Creus, Oriol Domingo
Professor: José Adrián Fonollosa