Anonymization

Project (T18693)

Karla Peña Ramírez

21/12/2022

An initiative of
the Netherlands
Red Cross

Removal of personal indentifiable information (PII) from textual data is one method to **ensure data privacy**.

We can do it **automatically and at scale** using AI (not perfectly, but consistent, reproducible, auditable and shareable).

Executive summary and recommendations

Two models were considered: English trained [main] and Spanish trained plus cased [alternative]. Their performance were **tested with four labelled datasets.**

- The **capitalization** of the entry names affects the model performance.
- The recovery rate of female (Spanish) names improves when the **model aggregation** (strategy to fuse (or not) tokens) is set to “first”.
- The **main model** mainly **fails** in the identification of names as **ORG** (both English and Spanish names, in particular the least frequent ones). Compound Spanish names (e.g. Ana María) are missed.
- The **alternative model outperforms the main model** (both English and Spanish). The large majority of the **missed names** are identified as **LOC**. A lot of people is named after geographical places!.
- The size of the testing dataset does not impact either of the model results.
- The **score output** value per tag can be used to narrow down the misidentifications.

Main model

English dataset: Reuters news stories between August 1996 and August 1997

English data	Articles	Sentences	Tokens
Training set	946	14,987	203,621
Development set	216	3,466	51,362
Test set	231	3,684	46,435

Train dataset: CoNLL-2003

Task-specific BERT model for name entity recognition (NER)

dslim/bert-base-NER

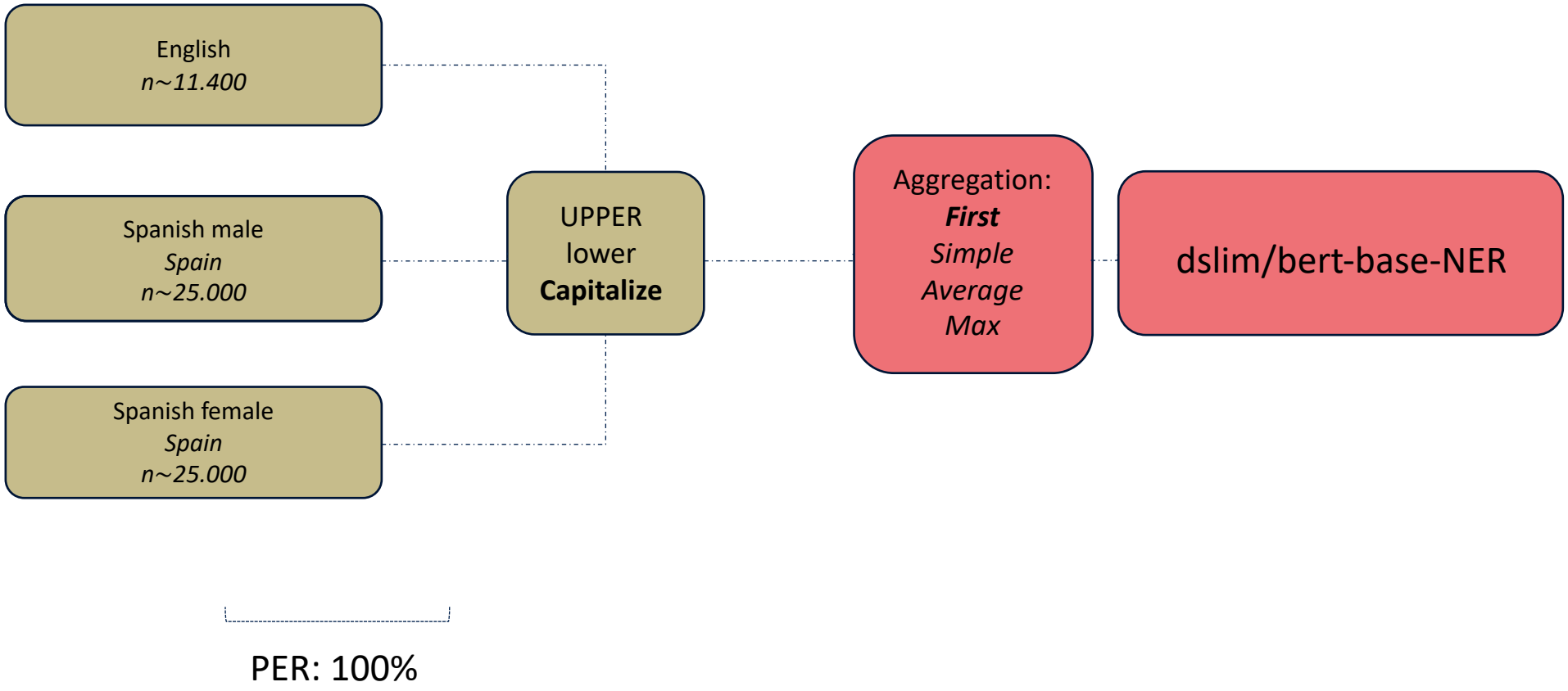
Output tags

English data	LOC	MISC	ORG	PER
Training set	7140	3438	6321	6600
Development set	1837	922	1341	1842
Test set	1668	702	1661	1617

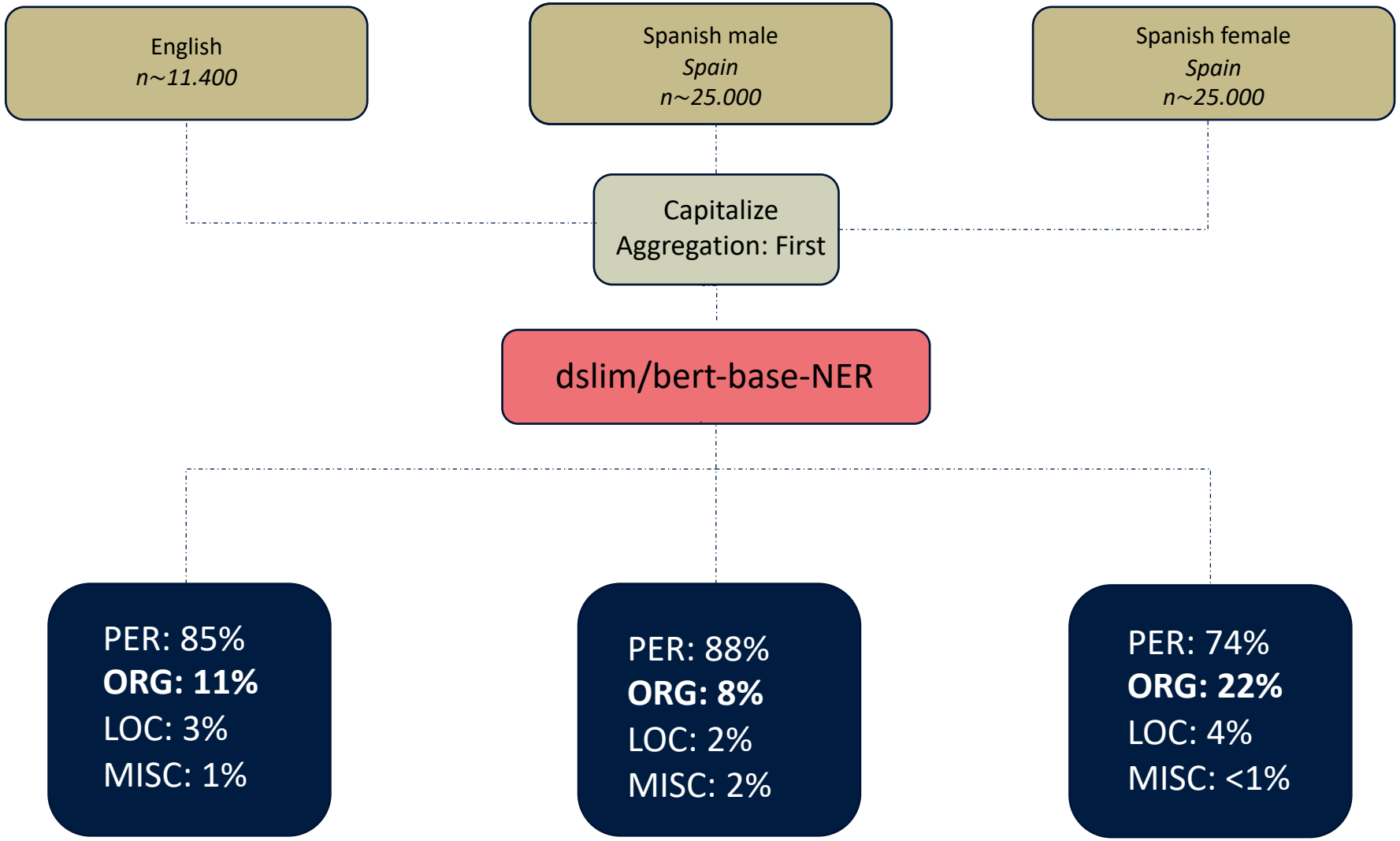
Location (LOC)
Organizations (ORG)
Person (PER)
Miscellaneous (MISC)

+ score (for each tag)

Main model implementation



Main model testing



Alternative model

Spanish dataset: EFE news stories on May 2000 (Tjong Kim Sang, 2002)

Train dataset: CoNLL-2002

Task-specific BERT model for name entity recognition (NER) **with fixed normalization**

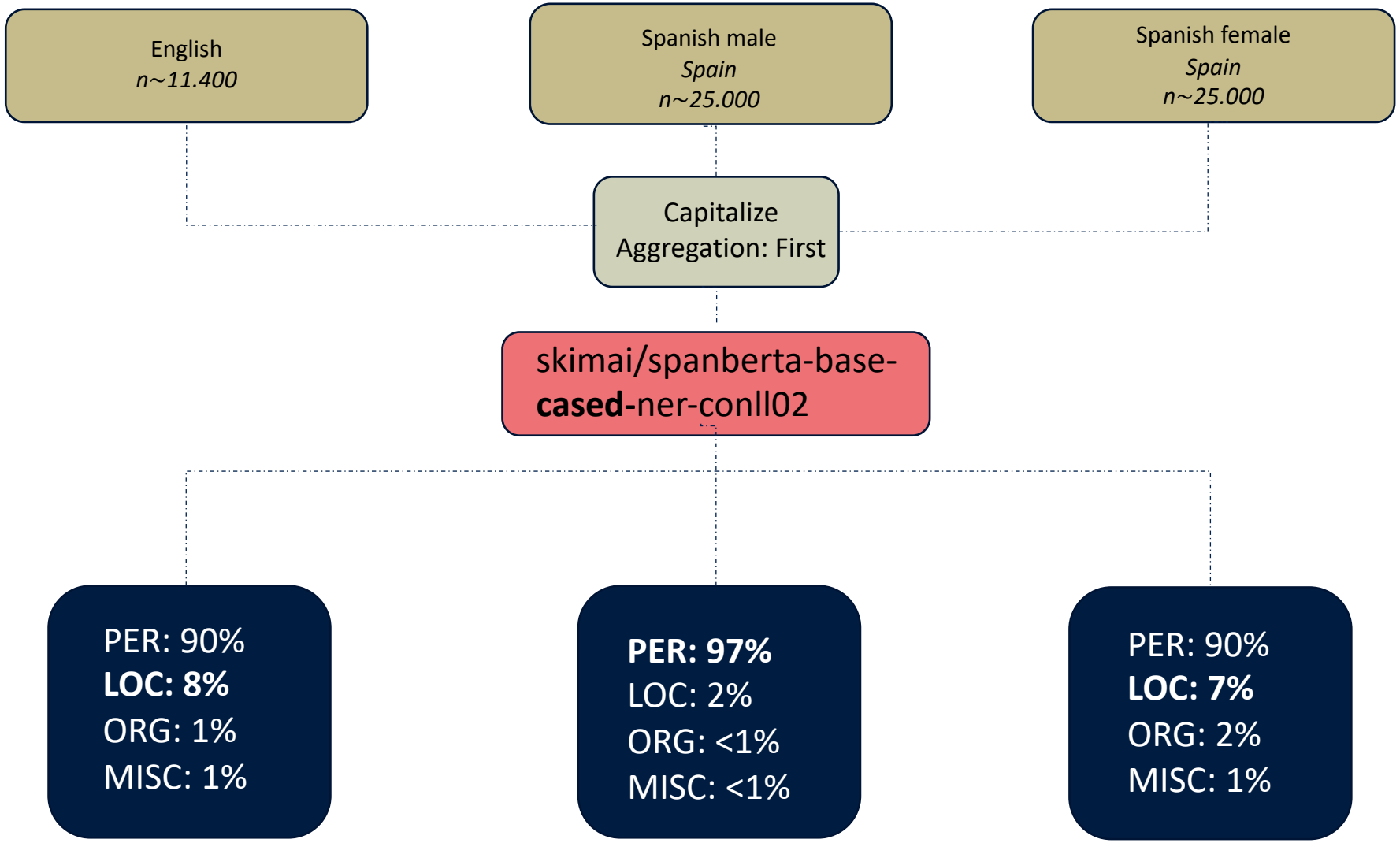
skimai/spanberta-base-**cased**-ner-conll02

Output tags

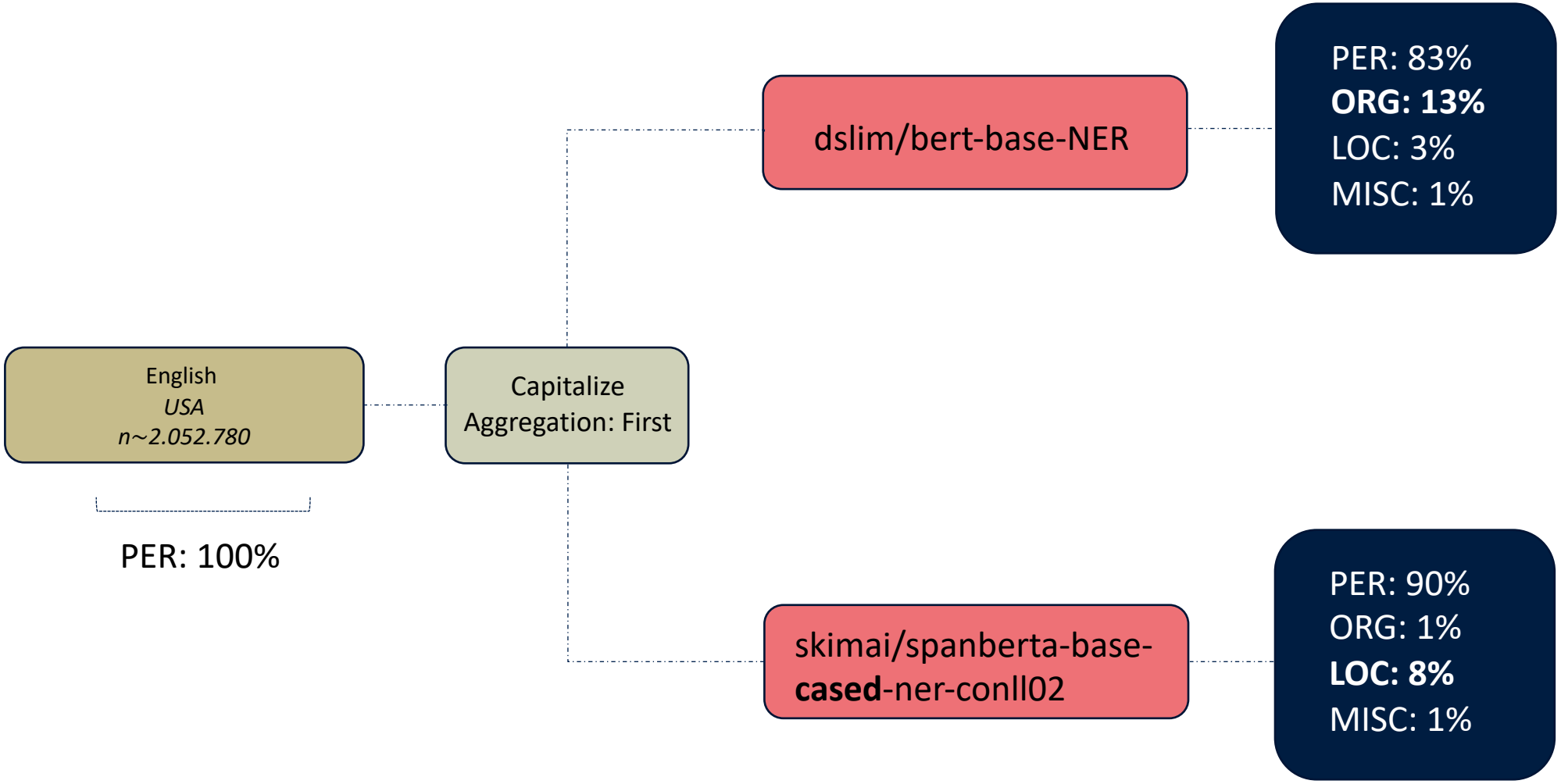
Location (LOC)
Organizations (ORG)
Person (PER)
Miscellaneous (MISC)

+ score (for each tag)

Alternative model testing



Large dataset (English)



A multilingual cased model trained with Wikipedia is avaialble:

https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip

https://github.com/google-research/bert/blob/master/multilingual.md

But…single-language models do a 3% better job than the Multilingual model.