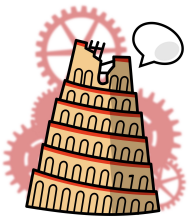
Can Cognate Prediction Be Modelled as a Low-Resource Machine Translation Task?

Clémentine Fourrier, Rachel Bawden, Benoît Sagot



INRIA (ALMAnaCH)

Aug 2021

Our task: Cognate prediction

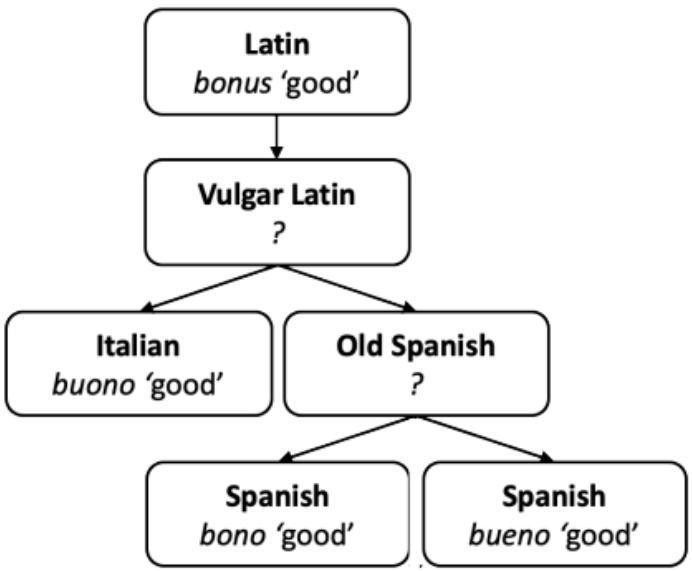


Figure 1: Cognates in our languages of interest

- **Cognates**
  Words in related languages evolved from the same ancestor

Our task: Cognate prediction

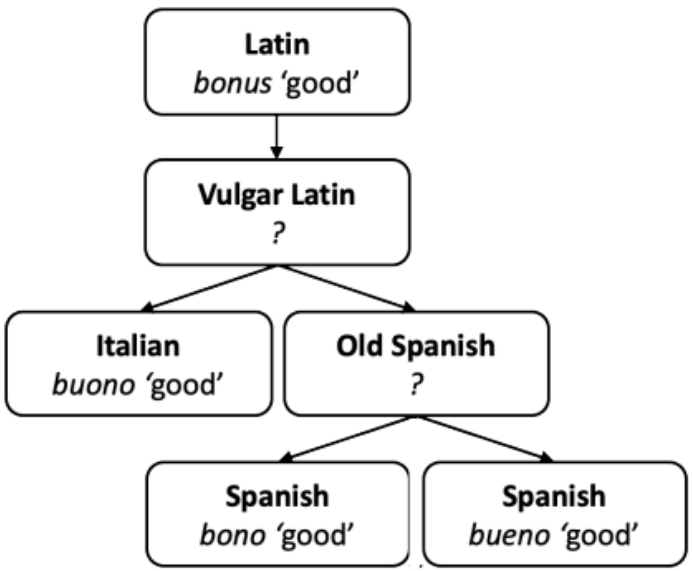


Figure 1: Cognates in our languages of interest

- **Cognates**
  Words in related languages evolved from the same ancestor
- **Cognate prediction**
  Producing likely cognates in related languages

Our task: Cognate prediction

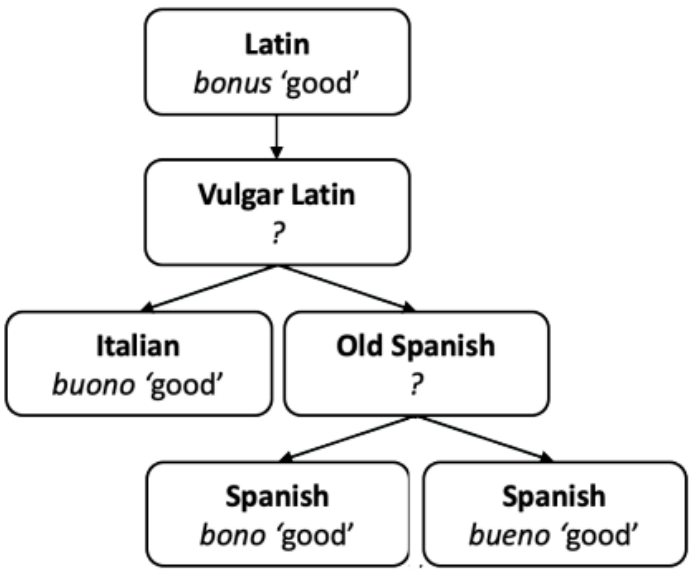


Figure 1: Cognates in our languages of interest

- **Cognates**
  Words in related languages evolved from the same ancestor
- **Cognate prediction**
  Producing likely cognates in related languages
  - Seq2seq correspondences
    $[b, u, o, n, o] \rightarrow [b, u, e, n, o]$
  - with very little data
  - ... similar to low resource MT?

Our task: Cognate prediction

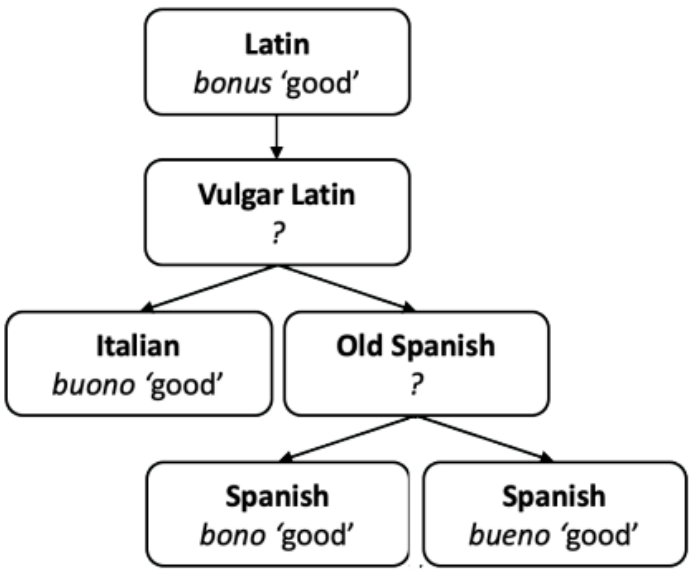


Figure 1: Cognates in our languages of interest

- **Cognates**
  Words in related languages evolved from the same ancestor
- **Cognate prediction**
  Producing likely cognates in related languages
  - Seq2seq correspondences
    $[b, u, o, n, o] \rightarrow [b, u, e, n, o]$
  - with very little data
  - ... similar to low resource MT?
- **Theoretical comparison**
  - Units and sample length
  - Modelled relations
  - Ambiguity management

Cognate prediction ↔ MT task?

Experiments

- Cognate data (ES-IT-LA)
- NMT (RNN, Transformers), SMT

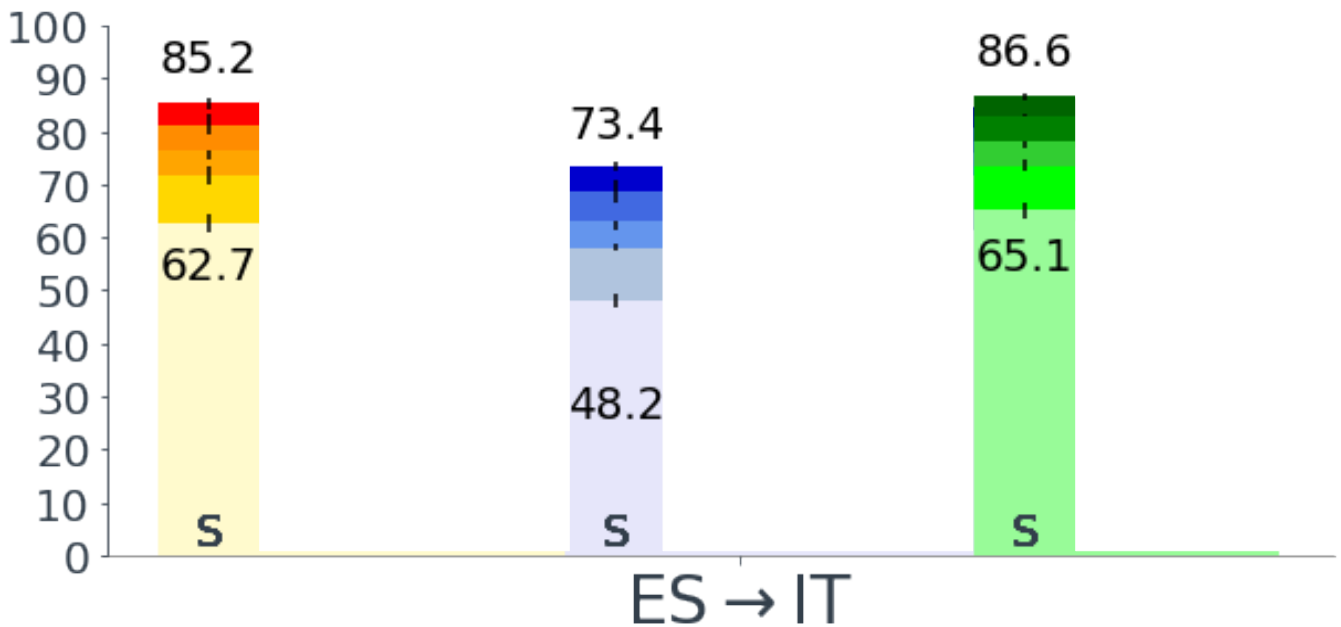


Figure 2: Results of our different setups for a language direction

Cognate prediction ↔ MT task?

Experiments

- Cognate data (ES-IT-LA)
- NMT (RNN, Transformers), SMT
- Data augmentation: pretrain

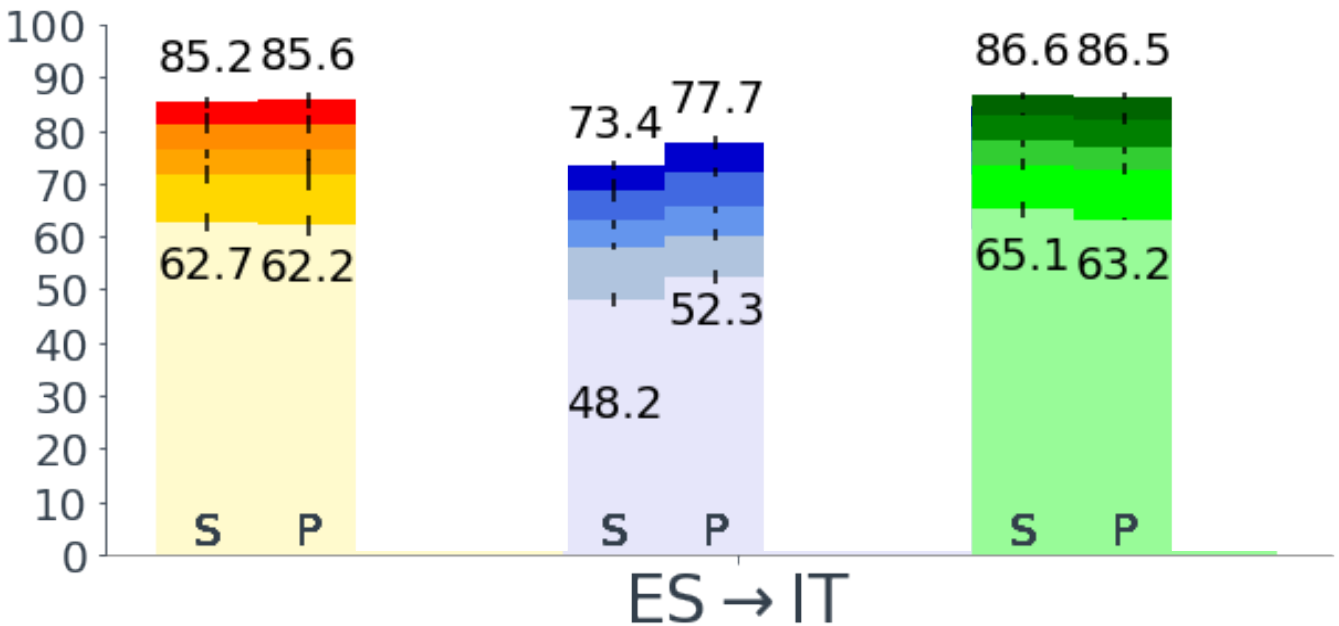


Figure 2: Results of our different setups for a language direction

Cognate prediction ↔ MT task?

Experiments

- Cognate data (ES-IT-LA)
- NMT (RNN, Transformers), SMT
- Data augmentation: pretrain, backtranslation

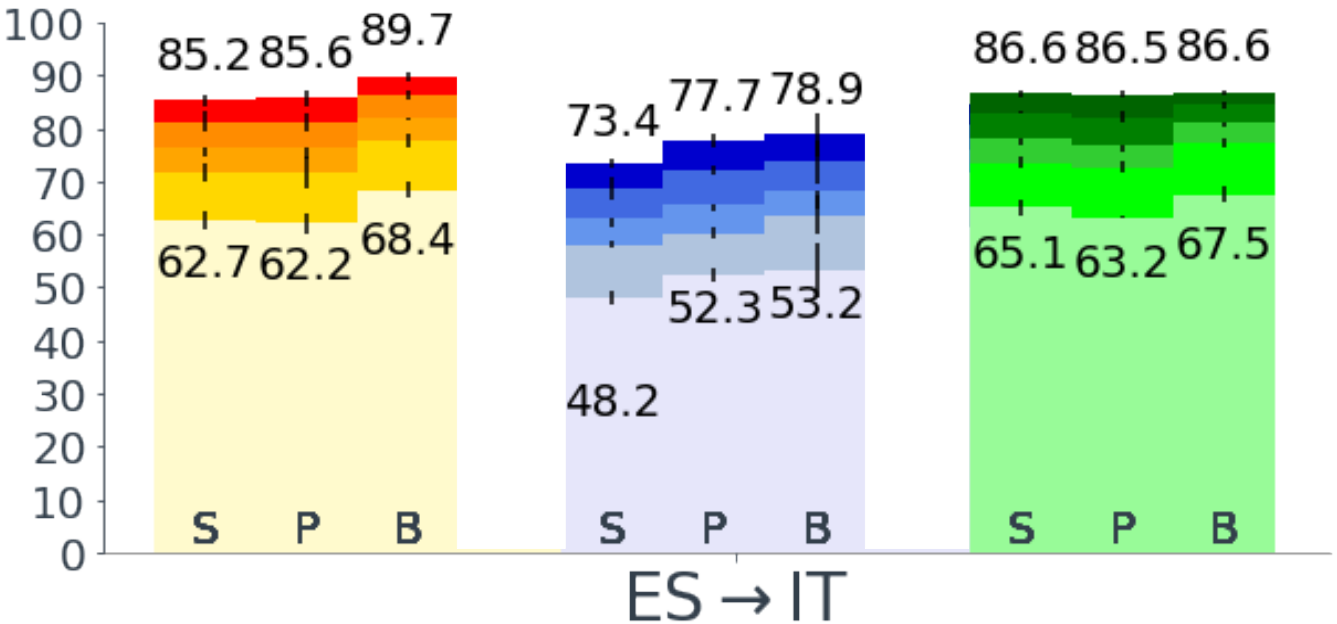


Figure 2: Results of our different setups for a language direction

Cognate prediction ↔ MT task?

Experiments

- Cognate data (ES-IT-LA)
- NMT (RNN, Transformers), SMT
- Data augmentation: pretrain, backtranslation, multilinguality

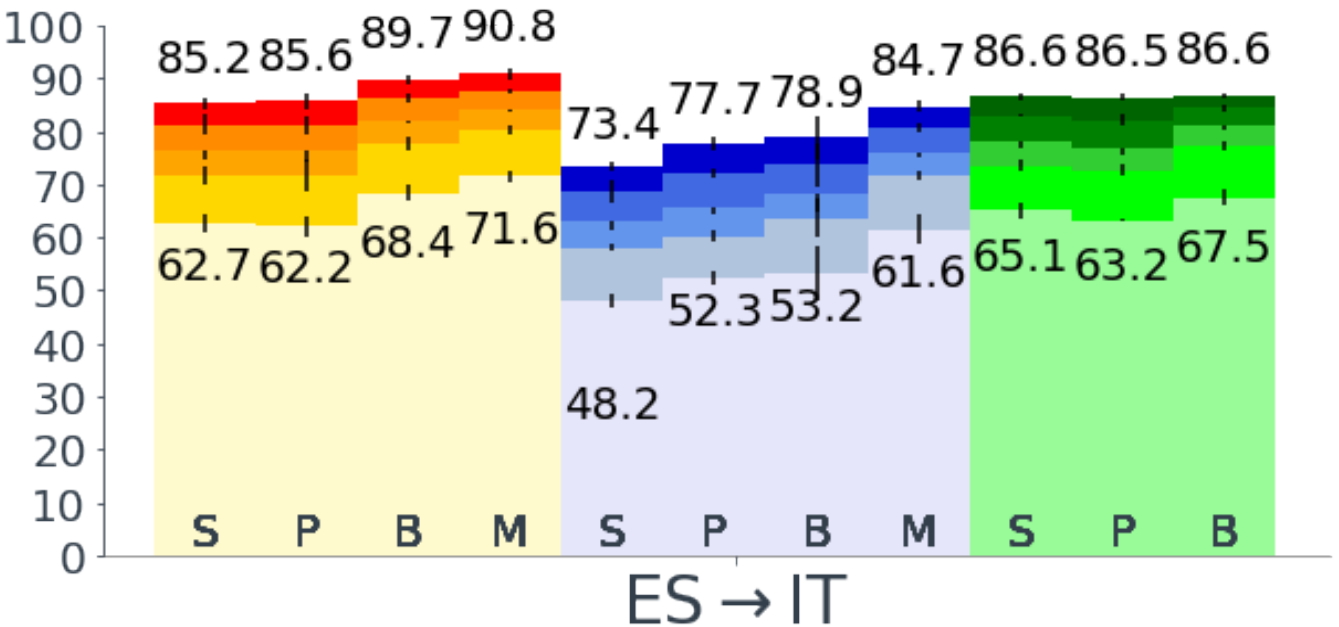


Figure 2: Results of our different setups for a language direction

Conclusion

About the methods

- SMT » NMT in low resource settings
- RNN » other methods when enough data
- Data augmentation:
  - through monolingual lexicons 😕
  - though multilinguality 🙂

Conclusion

About the methods

- SMT » NMT in low resource settings
- RNN » other methods when enough data
- Data augmentation:
  - through monolingual lexicons 😕
  - though multilinguality 🙂

About the tasks

- Because of ambiguity, need to predict n-best

Conclusion

About the methods

- SMT » NMT in low resource settings
- RNN » other methods when enough data
- Data augmentation:
  - through monolingual lexicons 😕
  - though multilinguality 🙂

About the tasks

- Because of ambiguity, need to predict n-best

→ Cognate prediction ↔ low-resource MT to an extent (needs to be studied with its specificities).

Thank you for listening, more info in the paper!
Poster session at LChange'21
Code and data at: github.com/clefourrier/CopperMT