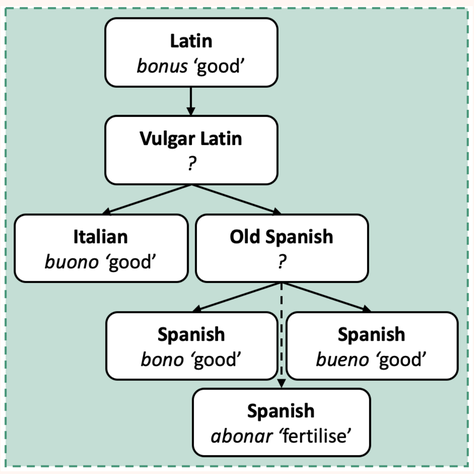
Can Cognate Prediction Be Modelled as a Low-Resource Machine Translation Task?

Clémentine Fourrier, Rachel Bawden and Benoît Sagot

Inria, Paris, France

What are cognates?

Cognates are evolutions, in related languages, of the same word from their common ancestor. (+ ancestor)



Latin *bonus* 'good' gave Italian *buono* 'id.', Spanish *bueno* 'id.' and Spanish *bono* 'id.', and they are all cognates.
Spanish *abonar* 'to fertilise', obtained by derivation, is related but not a cognate.

Goals

Cognate prediction produces likely cognates in related languages. Can be modelled as learning seq2seq correspondences from very little data.

→ Can cognate prediction be modelled as a low-resource machine translation task?

We study:

- their theoretical differences
- the efficiency of MT architectures for cognate prediction
- the transferability of data augmentation techniques (from low resource MT)

Theoretical differences

- Representation units
- Reordering, alignment
- Sample length
- Modelled relations
- Ambiguity management
- Impact of extra data for data augmentation

Data

Bilingual	LA–IT	LA–ES	ES–IT
#words	5,109	4,271	1,804
#phones	77,771	63,131	24,576
#Unique phones	34	39	38
Avg. word length	7.62	7.40	6.81
Monolingual	**ES**	**IT**	**LA**
#words	78,412	99,949	18,639
#phones	626,175	815,562	142,955
#Unique phones	38	40	29
Word length	7.98	8.24	7.67

Table 1:Dataset statistics for our lexicons.

Methods

Baseline (S):

- Neural MT:
  - Recurrent neural networks (orange - 1)
  - Transformers (blue - 2)
- Statistical MT (green - 3)

Data augmentation:

- Pre-training (P)
- Backtranslation (B)
- Multilingual neural translation (M)

Bilingual results

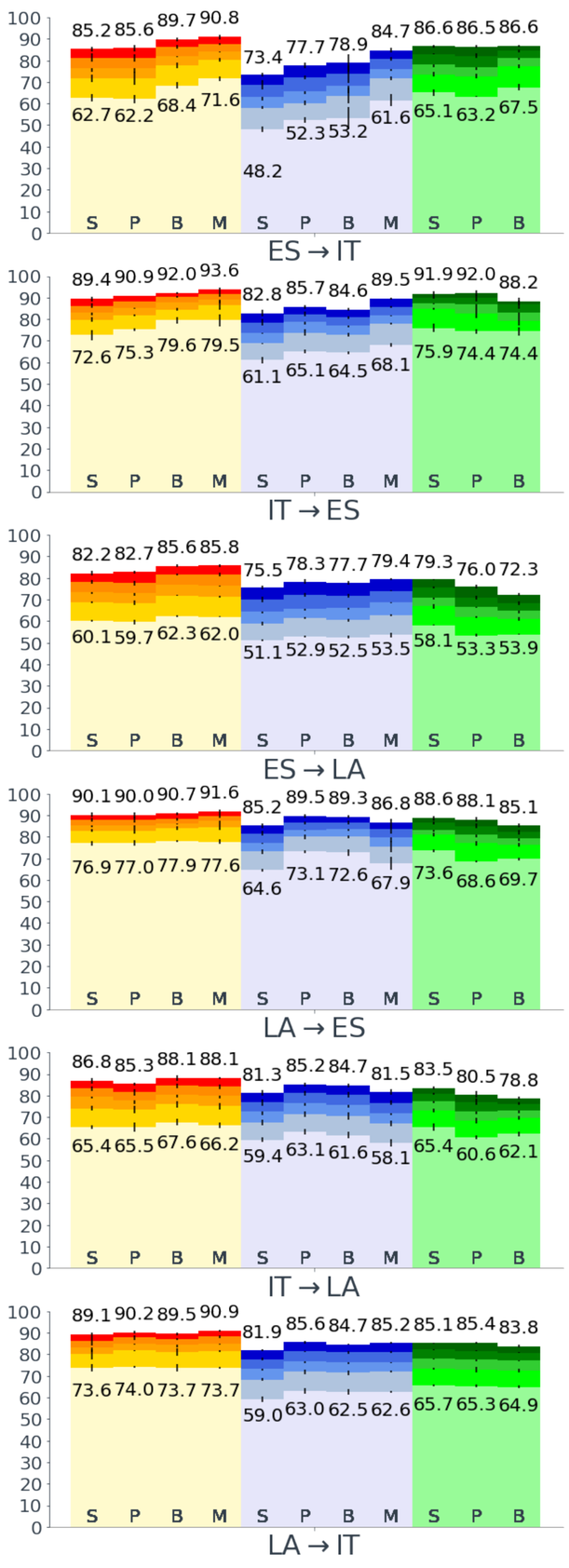


Figure 1:BLEU when comparing all methods

Multilingual results

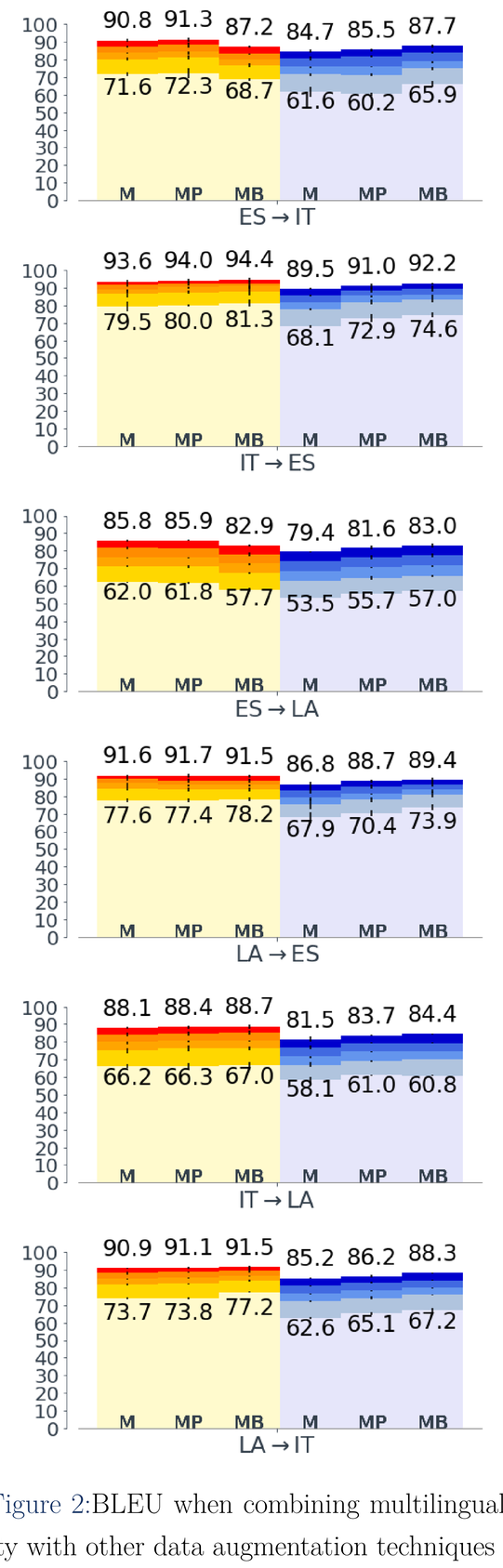


Figure 2:BLEU when combining multilinguality with other data augmentation techniques

Observations

- SMT > NMT for smallest datasets' baseline, RNN > otherwise
- Best data augmentation for RNNs: multilinguality
- Combining data augmentation methods is not very conclusive

Language choice for multilingual setups

Baseline	ES→IT	IT→ES
1000 pairs	53.9 ± 3.4	66.6 ± 4.2
Added data	**ES→IT**	**IT→ES**
Same language pair	62.5 ± 2.5	71.8 ± 1.7
Latin	57.1 ± 1.8	67.4 ± 3.3
French	58.5 ± 2.0	67.0 ± 2.8
Portuguese	58.8 ± 1.1	66.9 ± 2.9

Table 2:BLEU for different multilingual settings.

Best is to add data from current languages. If unavailable, add data from a parent language to source and target. If unavailable, add data from a language closest to the source.

Error analysis

BLEU prediction error repartition (for resp. SMT/RNN/Transformer):

- 84.6/81.4/79.5% : ambiguous sound correspondence
- 10.3/13.4/11.6% : rare sound correspondences
- 0.9/0.9/0.9% : data error
- 4.3/4.3/8.0% : model error

Conclusion

For our studied languages and methods: **Yes, as long as you take into account its specificities regarding ambiguity and n-best prediction.**

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011011459R1 made by GENCI. This work was partly funded by the last author's chair in the PRAIRIE institute funded by the French national agency ANR under the reference ANR-19-P3IA-0001.