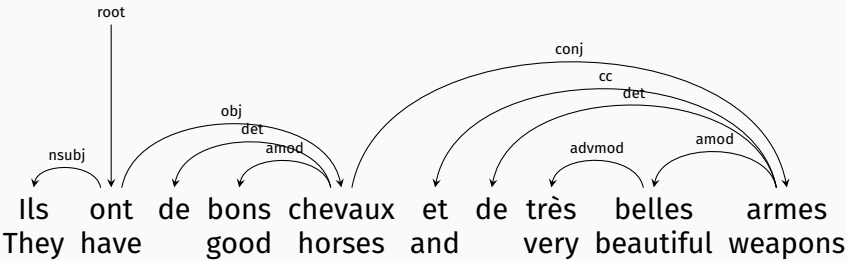
BERTrade: Using Contextual Embeddings to Parse Old French

Loïc Grobol, Mathilde Regnault, Pedro Ortiz Suarez, Benoît Sagot, Laurent Romary, Benoit Crabbé,

LREC 2022
Marseille, 2022-06

Motivation

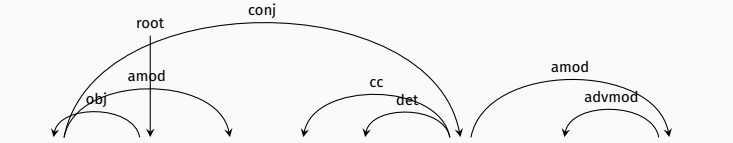
Le français contemporain



‘They have good horses and very beautiful weapons.’

- SVO
- No case system
- Overt subjects

Old French



‘They have good horses and very beautiful weapons.’

- Loose V2 with flexible word order
- Bicasual system
- Frequent null subject

What happened?

What happened?

- When did it happen?
- How did it happen?

→ We want empirical evidences, supported by corpora.

Corpora: SRCMF

Syntactic Reference Corpus of Medieval French a treebank of Old French from the 9th to the 13th century:

- 246 kwords, 23 ksentences

Corpora: SRCMF

Syntactic Reference Corpus of Medieval French a treebank of Old French from the 9th to the 13th century:

- 246 kwords, 23 ksentences
  - → Actually one of the largest treebanks of French in number of sentences!

Corpora: SRCMF

Syntactic Reference Corpus of Medieval French a treebank of Old French from the 9th to the 13th century:

- 246 kwords, 23 ksentences
  - → Actually one of the largest treebanks of French in number of sentences!
- Still too limited for a diachronic study of changes in French.

Corpora: SRCMF

Syntactic Reference Corpus of Medieval French a treebank of Old French from the 9th to the 13th century:

- 246 kwords, 23 ksentences
  - → Actually one of the largest treebanks of French in number of sentences!
- Still too limited for a diachronic study of changes in French.
- Not diverse enough, even on the target period.

Corpora: Profiterole

‘PRocessing Old French Instrumented TExts for the Representation Of Language Evolution’

- Extend SRCMF to all medieval French (9th–15th century) and 1 Mwords
- Annotating from scratch is very expensive

→ Let’s bootstrap it

  - Train parsers on SRCMF.
  - Use them to parse new data.
  - Correct the annotations.
  - Retrain the parsers.
  - Rince, repeat.

Corpora: Profiterole

‘PRocessing Old French Instrumented TExts for the Representation Of Language Evolution’

- Extend SRCMF to all medieval French (9th–15th century) and 1 Mwords
- Annotating from scratch is very expensive

→ Let’s bootstrap it

  - Train parsers on SRCMF.
  - Use them to parse new data.
  - Correct the annotations.
  - Retrain the parsers.
  - Rince, repeat.

HOPS

HOPS

A Honest Parser of Sentences (Grobol and Crabbé 2021), a graph parser

Graph parser

Le chat préfère le fromage

Graph parser

Le

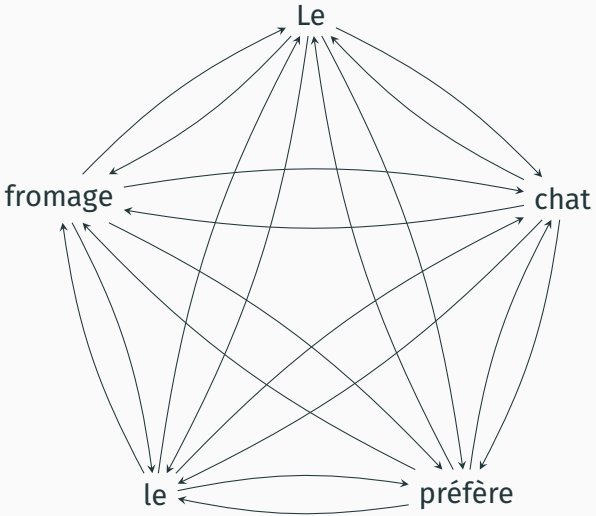
fromage

chat

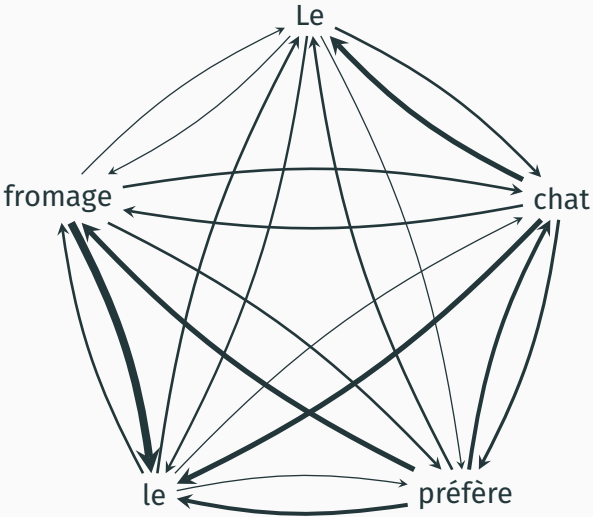
le

préfère

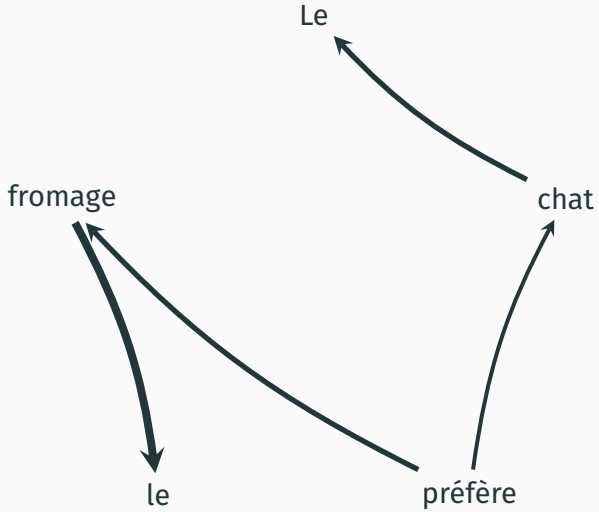
Graph parser



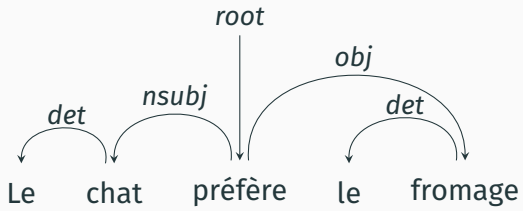
Graph parser



Graph parser



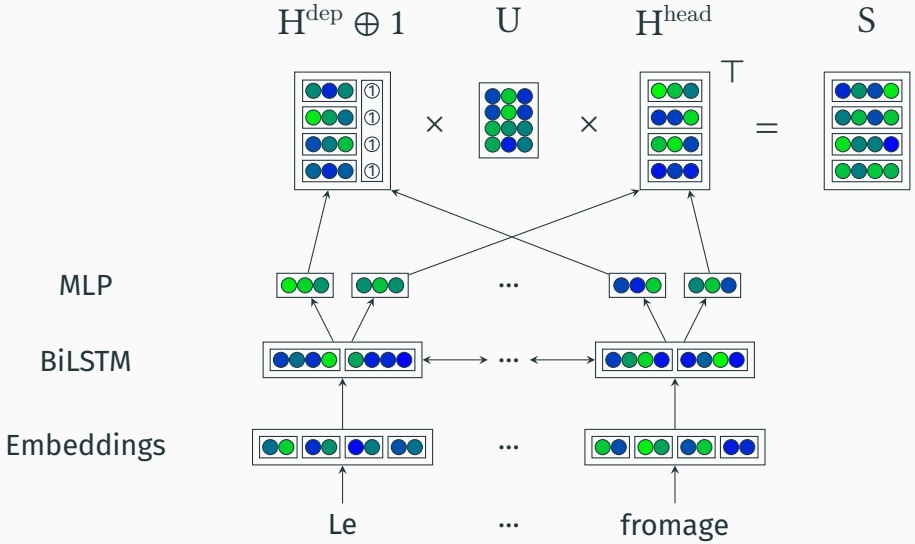
Graph parser



HOPS

A Honest Parser of Sentences (Grobol and Crabbé 2021), a graph parser using the biaffine architecture (Dozat and Manning 2017)

- A state-of-the-art neural parser:
  - Arbitrary word vector representations
  - A stack of BiLSTMs to build contextual representations
  - Shallow head/dependency label/POS prediction layers


$H^{dep} \oplus 1$
U
H^{head}
S
$\top$
MLP
BiLSTM
Embeddings
Le
...
fromage

HOPS

A Honest Parser of Sentences (Grobol and Crabbé 2021), a graph parser using the biaffine architecture (Dozat and Manning 2017)

- A state-of-the-art neural parser:
  - Arbitrary word vector representations
  - A stack of BiLSTMs to build contextual representations
  - Shallow head/dependency label/POS prediction layers
- Easy (≈) to implement
- Well known

Performances

Parsing on SRCMF with standard hyperparameters

- Decent
- Not good enough

Embeddings	UPOS	UAS	LAS
HOPS on SRCMF	93.51	87.60	81.54

Performances

Parsing on SRCMF with standard hyperparameters

- Decent
- Not good enough

Embeddings	UPOS	UAS	LAS
HOPS on SRCMF	93.51	87.60	81.54

We worked on contemporary French to help us understand why

Performances

Parsing on SRCMF with standard hyperparameters

- Decent
- Not good enough

Embeddings	UPOS	UAS	LAS
HOPS on SRCMF	93.51	87.60	81.54

We worked on contemporary French to help us understand why

→ State-of-the-art result
→ What **really** helps: having a BERT model

Performances

Parsing on SRCMF with standard hyperparameters

- Decent
- Not good enough

Embeddings	UPOS	UAS	LAS
HOPS on SRCMF	93.51	87.60	81.54

We worked on contemporary French to help us understand why

→ State-of-the-art result
→ What **really** helps: having a BERT model

But of course we don't have that for old French!

BERTrade

Sesame Street épisode 4192

BERT models are pretrained on a lot of data

- In the order of 1×10^8 words, 1×10^9 words
- Web scraping, Google books, Wikipedia

Of course, none of these really exist for Old French, so what to do?

BERT without data

Does using a BERT model without pretraining work?

BERT without data

Does using a BERT model without pretraining work?

Embeddings	UPOS	UAS	LAS
No BERT	93.51	87.60	81.54
Random BERT	93.17	86.97	80.71

BERT without data

Does using a BERT model without pretraining work?

Embeddings	UPOS	UAS	LAS
No BERT	93.51	87.60	81.54
Random BERT	93.17	86.97	80.71

No: it's worse than doing nothing.

Pretraining on other languages

Can we use a BERT model trained on contemporary French, the closest relative for which we do have such data?

Can we use a multilingual model?

Pretraining on other languages

Can we use a BERT model trained on contemporary French, the closest relative for which we do have such data?

Can we use a multilingual model?

Embeddings	UPOS	UAS	LAS
No BERT	93.51	87.60	81.54
FlauBERT	95.70	90.43	85.45
CamemBERT	95.86	91.15	86.31
mBERT	96.06	91.52	86.83

Pretraining on other languages

Can we use a BERT model trained on contemporary French, the closest relative for which we do have such data?

Can we use a multilingual model?

Embeddings	UPOS	UAS	LAS
No BERT	93.51	87.60	81.54
FlauBERT	95.70	90.43	85.45
CamemBERT	95.86	91.15	86.31
mBERT	96.06	91.52	86.83

- It **does** help a lot!
- The slight advantage for mBERT might be explained by more tolerance to variation.

Make do with what we have

There is no hope of gathering *BERT-like amounts of data: but we don’t have nothing.

Make do with what we have

There is no hope of gathering *BERT-like amounts of data: but we don't have nothing.

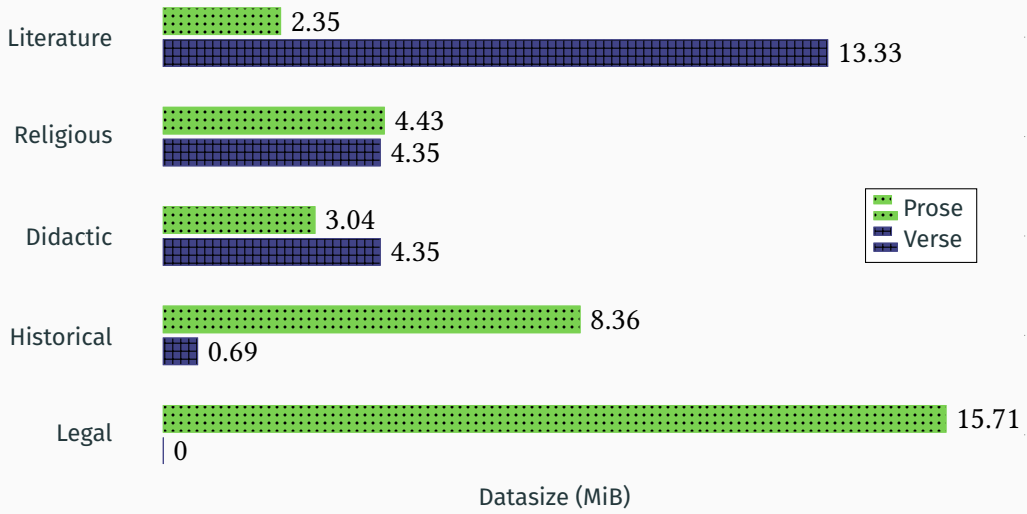
We compiled several small-to-medium scale corpora

- Nouveau Corpus d'Amsterdam
- Base de français mediéval
- Anglo-Norman database
- ...

Un corpus d'ancien français

Corpus	Size (MiB)	Size (Mwords)
BFM	20.7	3.91
AND	17.2	3.25
NCA	9.7	2.05
Chartes Douai	3.1	0.56
OpenMedFr	1.7	0.33
Geste	1.5	0.32
MCVF	1.4	0.26
Chartes Aube	0.2	0.04
Total	55.3	10.53

Un corpus d'ancien français



BERTrade from scratch

Is it enough?

- We are still several orders of magnitude below *BERT
- But it might be enough (Micheli et al. 2020)
  - → The secret sauce seems to be using deep but not very wide models

BERTrade from scratch

Is it enough?

- We are still several orders of magnitude below *BERT
- But it might be enough (Micheli et al. 2020)
  - → The secret sauce seems to be using deep but not very wide models

Let's try

Préentraîner BERTrade

We pretrain several BERT models on our raw corpus of OF, with varying sizes

Name	Layers	Embeddings	Heads	UPOS	UAS	LAS
mBERT	12	768	12	96.06	91.52	86.83
BERTrade-tiny	2	128	2	94.03	88.66	82.79
BERTrade-small	4	512	8	96.53	86.30	87.49
BERTrade-petit	12	256	4	**97.14**	**91.90**	**89.18**
BERTrade-medium	8	512	8	96.62	91.92	87.60
BERTrade-base	12	768	12	96.74	92.37	88.42

For all the serious configurations, this is better than using mBERT.

Posttraining

Can we go further?

- FlauBERT and CamemBERT have troubles adapting to OF
- Can we give them some help?

Posttraining

Can we go further?

- FlauBERT and CamemBERT have troubles adapting to OF
- Can we give them some help?

Would a crash course on raw Old French using our raw data help them?

Posttraining

It does!

Posttraining

It does!

Base model	UPOS	UAS	LAS
BERTrade-petit	97.14	92.95	89.18
BERTrade-mBERT	96.95	93.33	89.60
BERTrade-CamemBERT	97.16	93.75	90.06
BERTrade-FlauBERT	96.94	93.75	90.07

- This beats training from scratch.
- This time, monolingual models are better.

Test results

Compared to the (then) state of the art

Model	UPOS	UAS	LAS
Straka et al. (2019)	96.26	91.83	86.75
mBERT	96.19	92.03	87.52
BERTrade-petit	96.60	92.20	87.95
BERTrade-mBERT	97.11	93.86	90.37
BERTrade-FlauBERT	97.15	93.96	90.57
BERTrade-CamemBERT	97.29	94.36	90.90

Results on SRCMF test

Sesame Street épisode 4192

And now?

More data

- Using contemporary French helps
- But it is far from our target
- Can we use older historical French?

More data

- Using contemporary French helps
- But it is far from our target
- Can we use older historical French?

We can!

- Using a pre-1950 corpus of French extracted from the FranText base

→ 20 times bigger as our OF corpus

- For now the results are between our models trained from scratch and those adapted from contemporary French. The work goes on.
- Further plans: find a way to use the other Romance languages

An opportunistic use of raw data to parse Old French

- Collect as much in-domain data as possible helps, even if it is not much
- Adapt resources developed for contemporary French

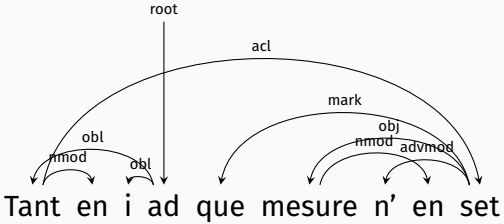
An opportunistic use of raw data to parse Old French

- Collect as much in-domain data as possible helps, even if it is not much
- Adapt resources developed for contemporary French

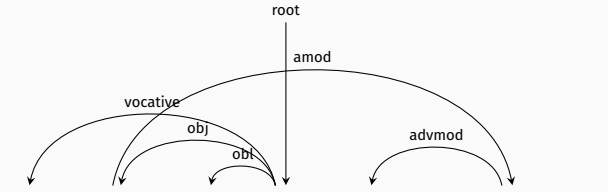
Ongoing work to adapt this to other historical languages: get in touch!

Appendix

Arbres remarquables



Arbres remarquables



‘Sire, I bring you very wonderful news’

References i

Dozat, Timothy and Christopher D. Manning (24th Apr. 2017). 'Deep Biaffine Attention for Neural Dependency Parsing'.
In: *Proceedings of the 5th International Conference on Learning Representations*. ICLR 2017 (Toulon, France). OpenReview.net.
URL: https://openreview.net/forum?id=Hk95PK9le.

Grobol, Loïc and Benoît Crabbé (June 2021). 'Analyse en dépendances du français avec des plongements contextualisés'.
In: *28e Conférence sur le Traitement Automatique des Langues Naturelles*. TALN 2021 (Lille, France).
Association pour le Traitement Automatique des Langues. URL: https://hal.archives-ouvertes.fr/hal-03223424.

Micheli, Vincent, Martin d'Hoffschmidt and François Fleuret (Nov. 2020).
'On the Importance of Pre-Training Data Volume for Compact Language Models'.
In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2020 (Online).
Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.emnlp-main.632.

Straka, Milan, Jana Straková and Jan Hajič (20th Aug. 2019).
'Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing'. arXiv: 1908.07448 [cs].
URL: http://arxiv.org/abs/1908.07448.

License