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
Ils ont de bons chevaux et de très belles armes
They have good horses and very beautiful weapons

‘They have good horses and very beautiful weapons.’

- SVO
- No case system
- Overt subjects
‘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*. 
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: 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!
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 TEtexts for the Representation Of Language Evolution’

- Extend SRCMF to all medieval French (9th–15th century) and 1Mwords
- 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.
  - Rinse, repeat.
Corpora: Profiterole

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

• Extend SRCMF to all medieval French (9th–15th century) and 1Mwords
• 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.
  • Rinse, repeat.
HOPS
A Honest Parser of Sentences (Grobol and Crabbé 2021), a graph parser
Le chat préfère le fromage
Le chat préfère le fromage
Le chat préfère le fromage.
Le chat préfère le fromage.
Le chat préfère le fromage.
Graph parser

Le chat préfère le fromage
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
\[ H^{\text{dep}} \oplus 1 \quad U \quad H^{\text{head}} \quad T = S \]

- **Embeddings**
- **BiLSTM**
- **MLP**

**Diagram:**
- Le
- ... fromage
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

<table>
<thead>
<tr>
<th>Embeddings</th>
<th>UPOS</th>
<th>UAS</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOPS on SRCMF</td>
<td>93.51</td>
<td>87.60</td>
<td>81.54</td>
</tr>
</tbody>
</table>
Performances

Parsing on SRCMF with standard hyperparameters

- Decent
- Not good enough

<table>
<thead>
<tr>
<th>Embeddings</th>
<th>UPOS</th>
<th>UAS</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOPS on SRCMF</td>
<td>93.51</td>
<td>87.60</td>
<td>81.54</td>
</tr>
</tbody>
</table>

We worked on contemporary French to help us understand why...
Performances

Parsing on SRCMF with standard hyperparameters

- Decent
- Not good enough

<table>
<thead>
<tr>
<th>Embeddings</th>
<th>UPOS</th>
<th>UAS</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOPS on SRCMF</td>
<td>93.51</td>
<td>87.60</td>
<td>81.54</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Embeddings</th>
<th>UPOS</th>
<th>UAS</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOPS on SRCMF</td>
<td>93.51</td>
<td>87.60</td>
<td>81.54</td>
</tr>
</tbody>
</table>

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
BERT models are pretrained on a lot of data

• In the order of $1 \times 10^8$ words, $1 \times 10^9$ words
• Web scraping, Google books, Wikipedia

Of course, none of these really exist for Old French, so what to do?
Does using a BERT model without pretraining work?
Does using a BERT model without pretraining work?

<table>
<thead>
<tr>
<th>Embeddings</th>
<th>UPOS</th>
<th>UAS</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>No BERT</td>
<td>93.51</td>
<td>87.60</td>
<td>81.54</td>
</tr>
<tr>
<td>Random BERT</td>
<td>93.17</td>
<td>86.97</td>
<td>80.71</td>
</tr>
</tbody>
</table>
Does using a BERT model without pretraining work?

<table>
<thead>
<tr>
<th>Embeddings</th>
<th>UPOS</th>
<th>UAS</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>No BERT</td>
<td>93.51</td>
<td>87.60</td>
<td>81.54</td>
</tr>
<tr>
<td>Random BERT</td>
<td>93.17</td>
<td>86.97</td>
<td>80.71</td>
</tr>
</tbody>
</table>

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?

<table>
<thead>
<tr>
<th>Embeddings</th>
<th>UPOS</th>
<th>UAS</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>No BERT</td>
<td>93.51</td>
<td>87.60</td>
<td>81.54</td>
</tr>
<tr>
<td>FlauBERT</td>
<td>95.70</td>
<td>90.43</td>
<td>85.45</td>
</tr>
<tr>
<td>CamemBERT</td>
<td>95.86</td>
<td>91.15</td>
<td>86.31</td>
</tr>
<tr>
<td>mBERT</td>
<td>96.06</td>
<td>91.52</td>
<td>86.83</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Embeddings</th>
<th>UPOS</th>
<th>UAS</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>No BERT</td>
<td>93.51</td>
<td>87.60</td>
<td>81.54</td>
</tr>
<tr>
<td>FlauBERT</td>
<td>95.70</td>
<td>90.43</td>
<td>85.45</td>
</tr>
<tr>
<td>CamemBERT</td>
<td>95.86</td>
<td>91.15</td>
<td>86.31</td>
</tr>
<tr>
<td>mBERT</td>
<td>96.06</td>
<td>91.52</td>
<td>86.83</td>
</tr>
</tbody>
</table>

• It **does** help a lot!
• The slight advantage for mBERT might be explained by more tolerance to variation.
There is no hope of gathering *BERT-like amounts of data: but we don’t have nothing.
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
• ...

Make do with what we have
<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size (MiB)</th>
<th>Size (Mwords)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFM</td>
<td>20.7</td>
<td>3.91</td>
</tr>
<tr>
<td>AND</td>
<td>17.2</td>
<td>3.25</td>
</tr>
<tr>
<td>NCA</td>
<td>9.7</td>
<td>2.05</td>
</tr>
<tr>
<td>Chartes Douai</td>
<td>3.1</td>
<td>0.56</td>
</tr>
<tr>
<td>OpenMedFr</td>
<td>1.7</td>
<td>0.33</td>
</tr>
<tr>
<td>Geste</td>
<td>1.5</td>
<td>0.32</td>
</tr>
<tr>
<td>MCVF</td>
<td>1.4</td>
<td>0.26</td>
</tr>
<tr>
<td>Chartes Aube</td>
<td>0.2</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>55.3</strong></td>
<td><strong>10.53</strong></td>
</tr>
</tbody>
</table>
Un corpus d’ancien français

<table>
<thead>
<tr>
<th>Genre</th>
<th>Prose</th>
<th>Verse</th>
<th>Datasize (MiB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature</td>
<td>2.35</td>
<td></td>
<td>13.33</td>
</tr>
<tr>
<td>Religious</td>
<td>4.43</td>
<td>4.35</td>
<td>8.36</td>
</tr>
<tr>
<td>Didactic</td>
<td>3.04</td>
<td>4.35</td>
<td></td>
</tr>
<tr>
<td>Historical</td>
<td></td>
<td></td>
<td>0.69</td>
</tr>
<tr>
<td>Legal</td>
<td></td>
<td></td>
<td>15.71</td>
</tr>
</tbody>
</table>
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
BERT Trade 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
We pretrain several BERT models on our raw corpus of OF, with varying sizes:

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

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?
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!

<table>
<thead>
<tr>
<th>Base model</th>
<th>UPOS</th>
<th>UAS</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERTrade-petit</td>
<td>97.14</td>
<td>92.95</td>
<td>89.18</td>
</tr>
<tr>
<td>BERTrade-mBERT</td>
<td>96.95</td>
<td>93.33</td>
<td>89.60</td>
</tr>
<tr>
<td>BERTrade-CamemBERT</td>
<td>97.16</td>
<td>93.75</td>
<td>90.06</td>
</tr>
<tr>
<td>BERTrade-FlauBERT</td>
<td>96.94</td>
<td>93.75</td>
<td>90.07</td>
</tr>
</tbody>
</table>

• This beats training from scratch.
• This time, monolingual models are better.
## Test results

Compared to the (then) state of the art

<table>
<thead>
<tr>
<th>Model</th>
<th>UPOS</th>
<th>UAS</th>
<th>LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Straka et al. (2019)</td>
<td>96.26</td>
<td>91.83</td>
<td>86.75</td>
</tr>
<tr>
<td>mBERT</td>
<td>96.19</td>
<td>92.03</td>
<td>87.52</td>
</tr>
<tr>
<td>BERTrade-petit</td>
<td>96.60</td>
<td>92.20</td>
<td>87.95</td>
</tr>
<tr>
<td>BERTrade-mBERT</td>
<td>97.11</td>
<td>93.86</td>
<td>90.37</td>
</tr>
<tr>
<td>BERTrade-FlauBERT</td>
<td>97.15</td>
<td>93.96</td>
<td>90.57</td>
</tr>
<tr>
<td>BERTrade-CamemBERT</td>
<td>97.29</td>
<td>94.36</td>
<td>90.90</td>
</tr>
</tbody>
</table>

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?

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
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

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

Tant en i ad que mesure n’ en set
Arbres remarquables

‘Sire, I bring you very wonderful news’

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.


This document is distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0)
(creativecommons.org/licenses/by/4.0)

© 2021, Loïc Grobol <loic.grobol@gmail.com>
http://www.llf.cnrs.fr/fr/Gens/Grobol