Automatic Normalisation of Modern French (i.e. from the 17th century)

LREC - 2022






Rachel Bawden
Jonathan Poinhos
Eleni Kogkitsidou
Philippe Gambette
Benoît Sagot
Simon Gabay

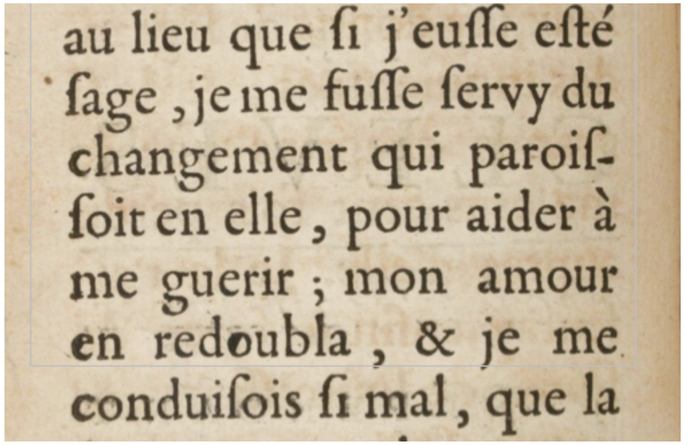
Inria
PR[AI]RIE PaRis Artificial Intelligence Research InstitutE
Université Gustave Eiffel
INSTITUT D'ÉLECTRONIQUE ET D'INFORMATIQUE GASPARD-MONGE
cnrs
UNIVERSITÉ DE GENÈVE

What is normalisation?

Application of a predefined convention to smooth out variation.
We choose contemporary French as the norm.

Why normalise modern French texts?

- Spelling conventions were not yet fixed, so there is a high degree of variability between texts
- The texts look a lot like contemporary French, but the differences make it hard to apply standard French tools to them

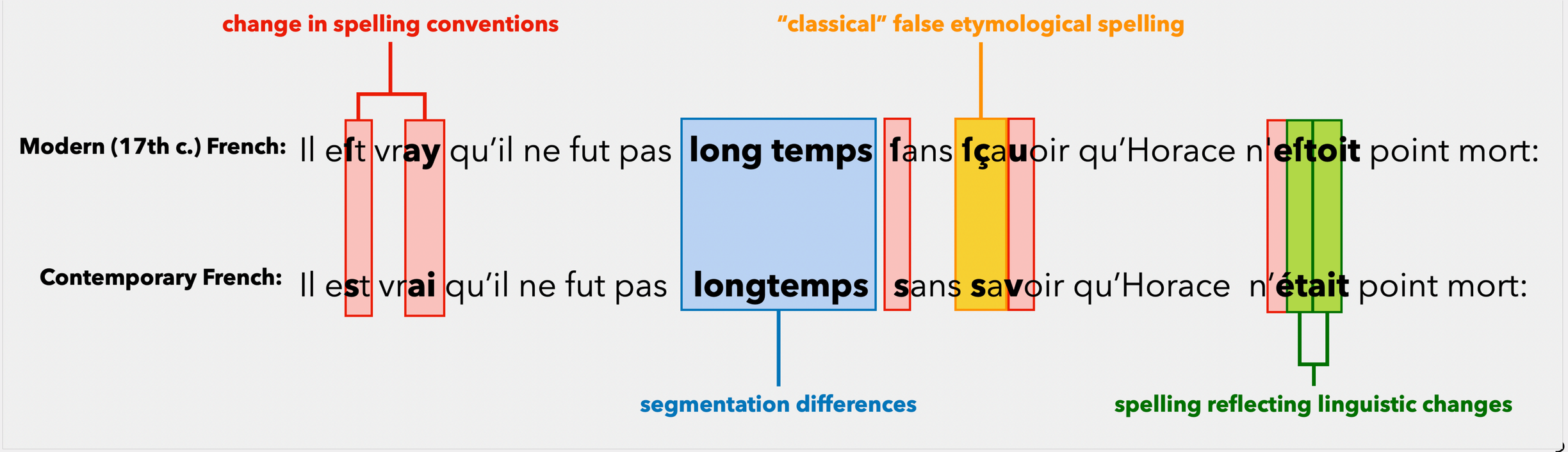

au lieu que ſi j'euſſe eſté
ſage, je me fuſſe ſervy du
changement qui paroiſ-
ſoit en elle, pour aider à
me guerir ; mon amour
en redoubla, & je me
conduiſois ſi mal, que la

	Accuracy (%)
Original	88.93
Normalised	91.10

PoS-tagging results when fine-tuning and testing French language model CamemBERT on 17th c. texts

Modern French vs. Contemporary French

- Some trivial changes (e.g. long s)
- Many non-trivial changes (segmentation differences, introduction of “classical” spellings, other changes indicative of language change, etc.)



Previous work on normalisation

- **Word lists, rules and edit-based approaches**
  - Replacing words by others depending on predefined lists or predefined correspondences (manual or automatic) (Baron and Rayson, 2009; Bollmann et al., 2011, Porta et al., 2013).
  - E.g. Levenshtein distance is a strong baseline (Pettersson et al., 2013)
- **MT approaches**
  - Most previous work has focused on character-based (learning correspondences of letters in a word), largely evaluated on normalisation of individual words (Vilar et al., 2007; Scherrer and Erjavec, 2013; Petters- son et al., 2013b; Domingo and Casacuberta, 2021)
  - Statistical MT (SMT) vs. neural MT (NMT): SMT can be superior if little data is available (Domingo and Casacuberta, 2018), but not always the case (Gabay and Barrault, 2020)

Contributions of this paper

1. New benchmark for the task: **parallel training data** for the normalisation of modern French into contemporary French

2. Development of **normalisation models** for Modern French: ruled-based, statistical, MT-inspired (statistical and neural)

3. **Evaluation metric** (symmetrised word accuracy) adapted to MT-inspired models and comparison of all models

FREEM$_{norm}$ dataset

https://freem-corpora.github.io/

- **Sentence-aligned parallel dataset (modern-contemporary French)**
- Genres: Caractères, comédie, tale, correspondence, law, fables, journalism, philosophy, poetry, novel, memoir novel, theology, tragedy, travel
  - Genres only available in test set: medicine, physics

		#unique tokens		#unique OOV tokens	
	#sents	**ModFr**	**Fr**	**ModFr**	**Fr**
Train	17.9k	264.3k	263.7k	-	-
Dev	2.4k	40.4k	40.3k	1.7k	1.3k
Test	5.7k	86.4k	86.2k	3.6k	2.5k

Normalisation methods

1. **Simple rule-based method (regular-expression-based)**
   - Manually written based on simple corpus statistics (some purely typographic and others lexical)
   - E.g. ſ → s, õ → om (if followed by m, b or p) and on (if not)
2. **Statistical alignment-based model (ABA)**
   - More details coming up!
3. **MT-based approaches:**
   - More details coming up!

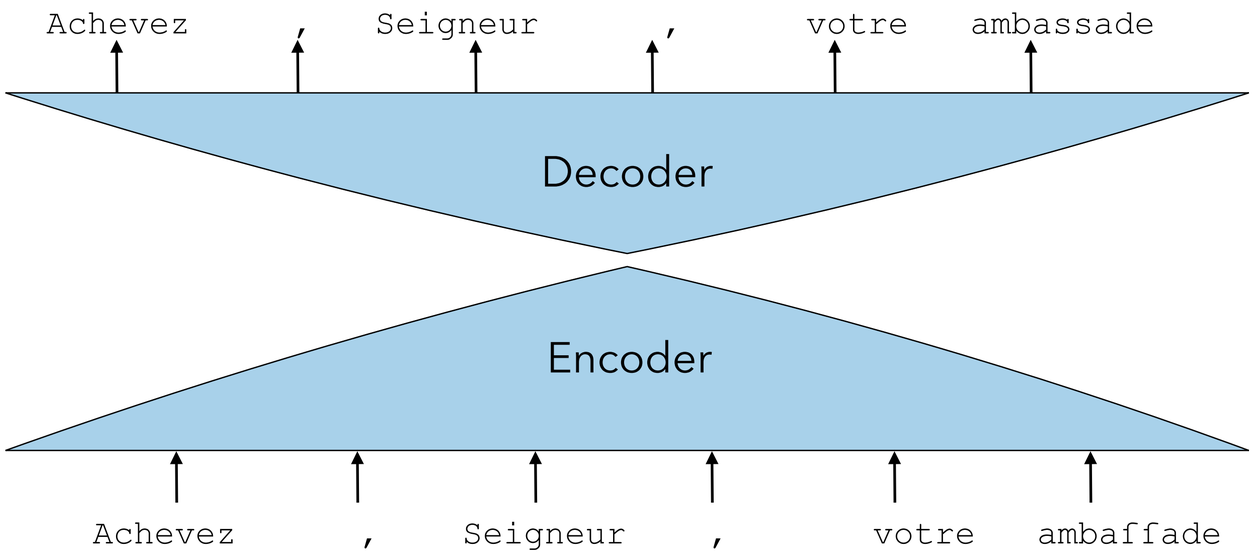
✳ **Optional lexicon-based post-processing step**

- To be applied after the other 3 methods
- Replace words that match modulo certain regular changes (e.g. accents, long s)

Alignment-based approach (ABA)

- Word-level translation rules learned from an aligned training corpus
- Character-level transformation rules manually designed by observing frequent transformations
- For each word not recognised as being contemporary French:
  - replace by the word in the word-level transformation rule if it exists
  - apply all possible combinations of character-level transformation rules, keep the first word existing in contemporary French, keep the original word otherwise

MT-style approaches



Advantages:

- Flexible word segmentation: allows for word merging or splitting
- Words are normalised in context (helpful in some cases and even necessary in others):

	Normalisation example 1	Normalisation example 2
nostre 'our'	quel malheur est le **nôtre** 'what woe is **ours**'	Les larmes sont trop peu pour pleurer **notre** mal 'The tears are too few to cry (for) **our** pain'
appellez 'call'	N'**appelez** point des yeux le Galant à votre aide 'Do not **call** the Galant for help with your eyes'	...Royaumes, par nous vulgairement **appelés** Siam '...kingdoms, **known** popularly by us as Siam'

MT-style approaches

- Statistical MT (SMT) - (1) Moses
- Neural MT (NMT) - Fairseq: (2) LSTM and (3) transformer

Extensive hyper-parameter searches:

- Subword segmentation (using sentencepiece and BPE):
  - Best subword segmentation with a vocabulary of 500 (SMT) and 1000 NMT
- Size of the networks (e.g. number of layers, embedding dimensions, etc.)
  - Best models were smaller than the base models used
- Learning rate and batch size

Evaluation

- Most commonly used metric = word/token-level accuracy
- In need of a reproducible implementation and one that is adapted to sentence-level normalisation:
  - Not necessarily a one-to-one token-level alignment
  - Hallucinations need to be penalised (risk of them being all associated with a single reference token)

Symmetrised accuracy:

Ref: Puisqu’ Achille combat , nous allons triompher

Hyp: Puisqu’ Achile combat , et oui nous allons triompher

According to reference tokenisation

Puisqu’	Achille	combat	,	nous	allons	triompher
Puisqu’	Achile	combat	,____et____oui	nous	allons	triompher

Accuracy = 6/7 = 0.86

Symmetrised acc = 0.82

According to hypothesis tokenisation

Puisqu’	Achile	combat	,	et	oui	nous	allons	triompher
Puisqu’	Achille	combat	,	▢	▢	nous	allons	triompher

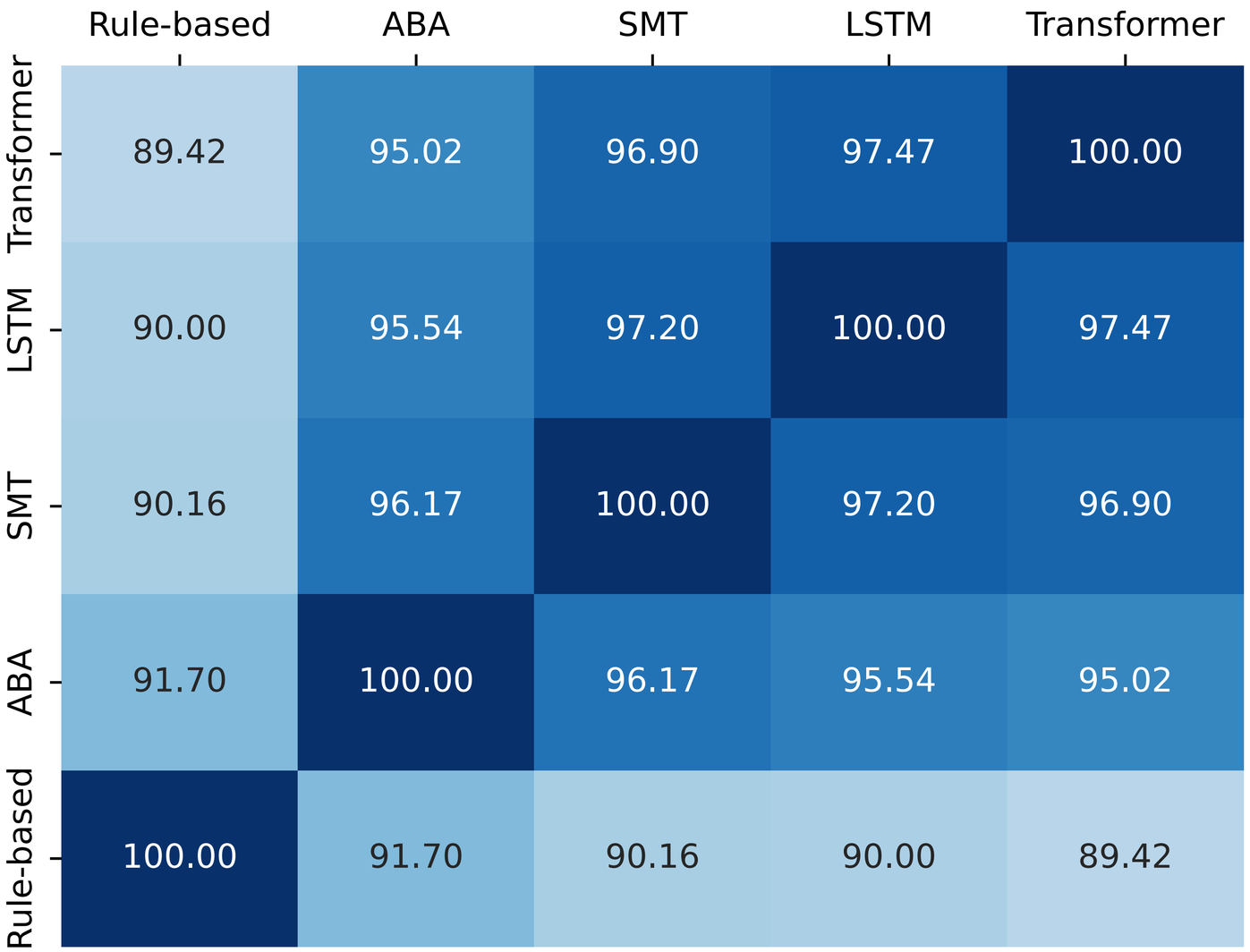
Accuracy = 7/9 = 0.78

Results

Method	**WordAcc (sym)**	**WordAcc (sym) OOV**
Identity	72.73	43.00
+ Le*fff*	86.12	64.84
Rule-based	89.05	60.22
+ Le*fff*	90.85	66.51
ABA	95.14	69.50
+ Le*fff*	95.44	73.54
SMT	**97.10±0.02**	76.64±0.18
+ Le*fff*	**97.24±0.02**	**78.37±0.20**
LSTM	96.14±0.08	**76.69±0.70**
+ Le*fff*	96.25±0.10	78.35±0.79
Transformer	95.89±0.08	75.73±0.38
+ Le*fff*	96.01±0.09	77.51±1.00

- Baselines already strong
- Best model = SMT
- Neural models do better on OOV words
- **Postprocessing (+ Le*fff*):**
  - Helps all methods
  - SMT+Le*fff* leads to best OOV scores

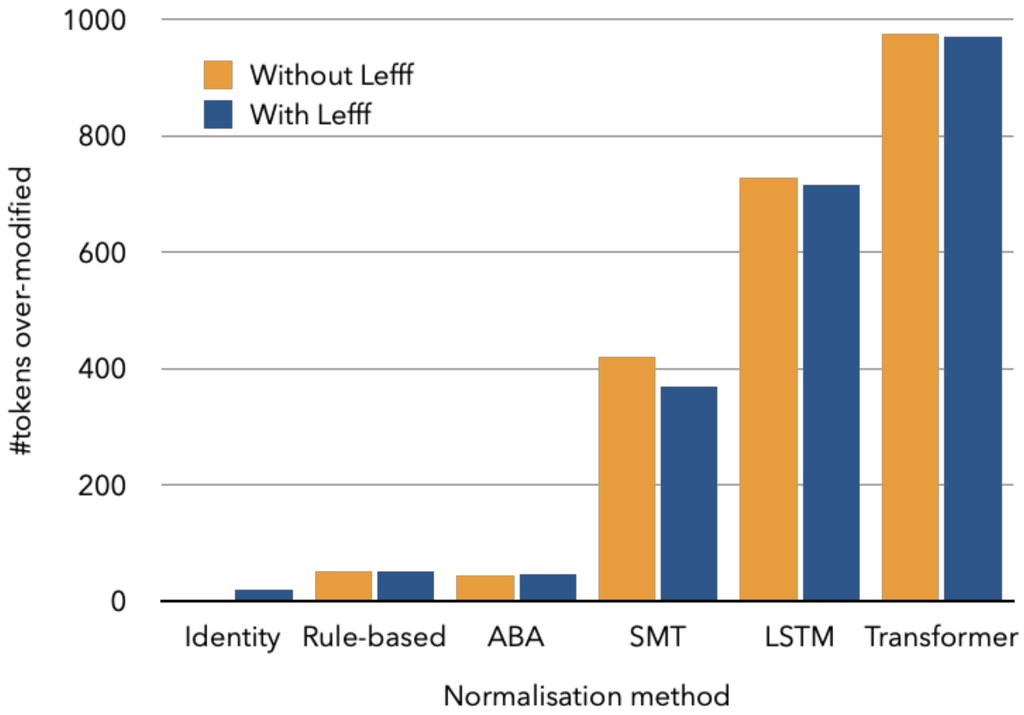
Similarity of the approaches



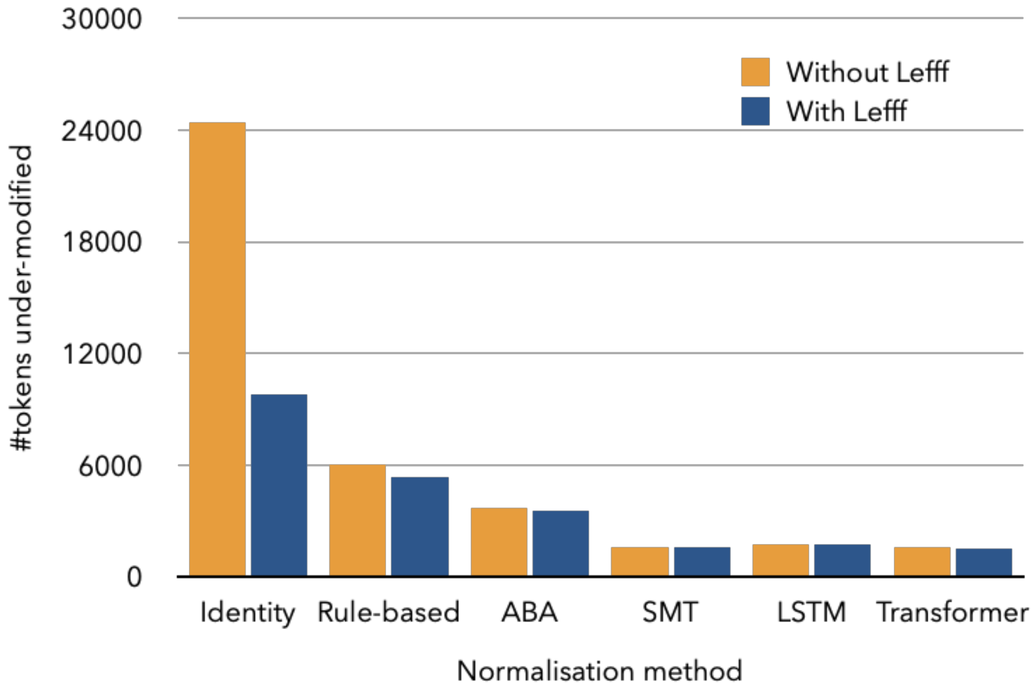
- Neural methods most similar (LSTM, Transformer)
- SMT is most similar to Transformer
- ABA most similar to rule-based

How zealous/conservative are the models?

Over-modifications



Under-modifications



What is better? This is actually task-dependent:

- As an aid for manual normalisation: conservative better
- For a down-stream annotation task (e.g. PoS tagging): zealous better

What sort of differences are there?

Comparison of the best rule-based approach (ABA+Le*fff*) and best MT approach (SMT+Le*fff*)

- **ABA is less robust to inadequacies in the training corpus**
  - E.g. succeeds with *auoient* → *avaient*, but not *avoient* → *avaient* (whereas SMT succeeds)
  - Lacks some rules (e.g. dealing double consonants)
- **SMT is in general more "creative":**
  - Some more creative errors (quite say to spot): *ma pẽʃée* 'pensée' -> *pmentsée*
  - Language model effect can be too strong (removes some determiners)
  - But handles ambiguity better (because it is contextual)
    - ABA+Lefff: *Car enfin n'attends pas que mes feux* ***redoublez*****,**
    - SMT+Lefff: *Car enfin n'attends pas que mes feux* ***redoublés*****,**
- The approaches appear to be complementary - potential for combining the two!

Conclusion and perspectives

- New benchmark for the normalisation of Modern French into contemporary French (dataset, baselines and state-of-the-art models)
- Comparison of different approaches (rule-based and MT-inspired) with different advantages
  - Potential for combining them
  - Further experiments required to test which model is best in different scenarios (i.e. to aid manual normalisation or for downstream tasks).
- Aim to facilitate and encourage research on Modern French

Thank you very much!

Our code and models are freely available:
https://github.com/rbawden/ModFr-Norm

Further information on the FreEm project page:
https://freem-corpora.github.io