Evaluation of methods and tools for automatic lemmatization in Old French

Cristina G. Holgado[1], Alexei Lavrentev[2], Mathieu Constant[3]

(1) Université de Strasbourg (Strasbourg, France)
(2) CNRS, ENS de Lyon, IHRIM (Lyon, France)
(3) Université de Lorraine, CNRS, ATILF (Nancy, France)

In the framework of the ANR project « PROFITEROLE »

TALN 2021 – 28 Juin - 2 Juillet

Context

- Projet **Profiterole** (ANR-16-CE38-0010)
- **Different texts** : date (main factor), dialect/scripta, domain-genre, form
- State of the art of automatic lemmatization :
  - Rule systems and lexicons :
    - TreeTagger (Schmid, 1994)
    - LGeRM (Lemmas, Graphs and Morphological Rules) (Souvay, 2009)
  - Supervised learning :
    - UDPipe (Straka et al., 2016) — *R (UDPipe v0.8.3)*
    - Pie (Manjavacas et al., 2019) — *(v0.8.5)*
- Special focus on lemmatization for unknown forms

Lemmatizing medieval French: what challenges?

- Lack of graphic standardisation: **form** → several spellings, higher ambiguity

  « *a* » → VERB/PREPOSITION
- Morphological complexity & dialectal variation
- Different dictionaries: sometimes different form entries for the same lexeme (some privilege modernised forms, others medieval ones)

  → **Lemma standardization**

  Standardized corpus: **DMF** (98,54%), **DECT** (1,04%), **BFM** (0,22%), **TL** (0,18%), **GDF** (0,02%)
- Lexical resources and works on NLP are not very advanced for this language

Tools

Rules and lexicons

TreeTagger

- Morphosyntactic tagging: decision trees learned from annotated corpus
- Lemmatization: based on the POS tag, lemmas are searched in the lexicon of the training corpus
- Unknown form <*-no-unknown*> : lemma = form
- External lexicon (not used)

LGeRM

- Developed for the Middle French
- Browses through its own enriched lexicon: Dictionary of Middle French (DMF)
- Complex system of rules for unknown forms of the lexicon
- Uses TreeTagger to filter ambiguity cases from predicted POS

Tools

Supervised learning (context)

UDPipe

- Designed for "universal" use
- Lemmatization rules learned from triplets (form, POS, lemma)
- Performs disambiguation

Pie

- Designed for non-stabilised historical languages
- Neural encoder-decoder model
- Independent POS/lemma
- Joint learning for prediction of next and previous words

Corpus

- Part of the **Base de Français Médiéval** (*Old French Database*)
- Annotated corpus composed of **431,144** tagged and lemmatized forms
- Two sources :
    - A predominant one: a single author (Chrétien de Troyes), same genre. Lemmatized at ATILF in the framework of the project DÉCT (Souvay & Kunstmann, 2008) → **254 000 forms**
    - Other: Diverse genres. Lemmatised in the framework of the BFM (IHRIM) → **177 000 forms**

Texts of the test sets

Test	Date	Dialect	Genre	Tokens	unknown
1	late 12e s.	*champenois*	novel	254 094	11,4%
2	late 12e s.	*champenois*	novel	47 965	2,6%
3	mid 11e s.	*normand*	hagiography	5530	13,9%
4	early 12e s.	*normand*	epic	35 312	15,3%
5	mid 12e s.	*anglo-normand*	chronicle	18 021	18,8%
6	early 14e s.	*no dialectical traits*	chronicle	11 035	12,7%
7	late 13e s.	*no dialectical traits*	hagiography	22 769	8,81%
8	early 11e s.	*franco-occitan*	hagiography	5092	31,8%
9	mid 13e s.	*hainaut*	charter	10 492	16,3%
10	late 14e s.	*no dialectical traits*	register	11 981	19,9%

TABLE 1: Characteristics of the texts in every test set

Overview of the results

LGeRM

- Rich lexicon
- Low precision for older lemmas

TreeTagger

- Performance affected by training corpus size
- Poor prediction for unknown forms (except form = lemma; proper nouns, infinitives)

Pie / UDPipe

- Better performance for unknown forms
- Need more samples in some categories to generalize better

TreeTagger	LGeRM	UDPipe	Pie
All forms*			
0,74	**0,83**	0,66	0,66
Unknown forms			
0,12	0,68**	**0,14**	**0,23**

Table 2. Mean precision (micro) of all the lemmas and unknown forms

* Punctuation excluded, ** Most of the forms are found in the lexicon

Lemmatization by POS

All forms

Cat.	Tokens	%	m.inc.	%	TreeTagger		LGeRM		UDPipe		Pie	
					tout	inc.	tout	inc.	tout	inc.	tout	inc.
ADJ	14 773	4,03	2680	5,60	0,73	0,10	0,83	0,67	0,65	0,11	0,55	0,18
ADV	39 535	10,78	2435	5,09	0,71	0,10	0,63	0,75	0,81	0,13	0,62	0,18
CON	37 233	10,15	44	0,09	0,82	0,00	0,94	0,37	0,94	0,02	0,77	0,44
DET	35 853	9,77	812	1,70	0,68	0,06	0,81	0,55	0,72	0,05	0,65	0,15
Ncom	53 989	14,72	12 649	26,43	0,68	0,12	0,71	0,68	0,51	0,14	0,50	0,20
Npro	9268	2,53	6058	12,66	0,54	0,40	0,43	0,47	0,34	0,30	0,26	0,03
PRE	34 309	9,35	667	1,39	0,66	0,08	0,80	0,60	0,77	0,18	0,59	0,12
PRO	60 870	16,59	770	1,61	0,72	0,01	0,75	0,60	0,67	0,06	0,57	0,15
VER	80 522	21,95	21 413	44,74	0,62	0,02	0,84	0,80	0,57	0,13	0,60	0,36
Total	366 882		47 859									

TABLE 3: Precision (micro) by POS for all forms

Lemmatization by POS

Unknown forms

Cat.	Tokens	%	m.inc.	%	TreeTagger		LGeRM		UDPipe		Pie	
					tout	inc.	tout	inc.	tout	inc.	tout	inc.
ADJ	14 773	4,03	2680	5,60	0,73	0,10	0,83	0,67	0,65	0,11	0,55	0,18
ADV	39 535	10,78	2435	5,09	0,71	0,10	0,63	0,75	0,81	0,13	0,62	0,18
CON	37 233	10,15	44	0,09	0,82	0,00	0,94	0,37	0,94	0,02	0,77	0,44
DET	35 853	9,77	812	1,70	0,68	0,06	0,81	0,55	0,72	0,05	0,65	0,15
Ncom	53 989	14,72	12 649	26,43	0,68	0,12	0,71	0,68	0,51	0,14	0,50	0,20
Npro	9268	2,53	6058	12,66	0,54	0,40	0,43	0,47	0,34	0,30	0,26	0,03
PRE	34 309	9,35	667	1,39	0,66	0,08	0,80	0,60	0,77	0,18	0,59	0,12
PRO	60 870	16,59	770	1,61	0,72	0,01	0,75	0,60	0,67	0,06	0,57	0,15
VER	80 522	21,95	21 413	44,74	0,62	0,02	0,84	0,80	0,57	0,13	0,60	0,36
Total	366 882		47 859									

TABLE 4: Precision (micro) by POS for unknown forms

Common lemmatization errors

UDPipe		
form	**gold**	**predicted**
amfant	**NOMcom/**enfant	**VERppa/**amfer
oultre	**ADVgen/**outre	**VERcjg/**oultre
ycelle	**DETdem/**cil	**ADJqua/**yceux

Pie		
form	**gold**	**predicted**
Berthier	**NOMpro/**Berthier	**VERinf/**Berter
amfant	**NOMcom/**enfant	**NOMcom/**amprendre
ycelle	**DETdem/**cil	**DETdem/**iceller

TABLE 4: Samples of lemmatization errors

Conclusion

- Lexicons and rule systems: they benefit from rich morphological lexicons
- Supervised methods: good generative ability for unknown forms
- Improved representation of periods and dialects of Old French in the training corpus
- Use of the lexicon and rules of LGeRM in combination with more recent lemmatizers

References

- Manjavacas, Enrique, Ákos Kádár, et Mike Kestemont. 2019. « Improving Lemmatization of Non-Standard Languages with Joint Learning ». In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 1493-1503. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1153.
- Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees.
- Straka, Milan, Jan Hajic, et Jana Strakova. 2016. « UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing », 8.
- Souvay, Gilles, et Jean-Marie Pierrel. 2009. « LGeRM Lemmatisation des mots en Moyen Français ». Traitement Automatique des Langues 50 (2): 21.
- Kunstmann, Pierre, et Gilles Souvay. 2007. « DÉCT : Dictionnaire Électronique de Chrétien de Troyes ». In CILPR 2007 Congrès International de Linguistique et de Philologie Romane, xxx. Innsbruck, Austria. https://hal.archives-ouvertes.fr/hal-00418939.