Neural Lemmatization of Multiword Expressions

Marine Schmitt and Mathieu Constant
Université de Lorraine & CNRS, ATILF

Joint Workshop on Multiword Expressions and WordNet
August 2, 2019

Lemmatization of Multiword Expressions I

Lemmatization

- Task : given a form occurring in a text, automatically find its base form
- For simple words, problem (almost) solved
- What about multiword expressions ?

Example

John **spilled the beans** on government corruption

- form = spilled the beans
- lemma = spill the beans

Lemmatization of Multiword Expressions II

Motivations

- Natural next step after MWE identification and extraction
  $\{$*spilling the beans*, *spilled the beans*, ...$\} \rightarrow$ *spill the beans*
- Useful for MWE linking, especially in morphologically-rich languages

This talk

- Preliminary experiments on different languages
- French as pilot language

What is difficult with MWE lemmatization ?

Detect (in)variability

Form :	*pulled*	***strings***
Lemma :	pull	strings

Handle agreement

cartes	bleues		carte	bleue	
cards.NOUN.FEM.PL	blue.ADJ.FEM.PL	$\rightarrow$	card.NOUN.FEM.SG	blue.ADJ.FEM.SG	
"credit cards"			"credit card"		

Handle ambiguity. Fortunately...

- Ambiguity is very rare (in our datasets)
- Interest for unknown MWEs only

Related work

- Finite-state rule-based morphology analysis (Oflazer and Kuruoz, 1994 ; Oflazer et al., 2004)
- Rule-based approaches relying on dictionaries (Stankovic et al., 2016 ; Marcinczuk, 2017)
- Statistical tagging + dictionary lookup (Radziszewski, 2013)
- Recent related shared tasks including lemmatization of multiword noun phrases and named entities in Slavic languages : PolEval 2019, BSNLP 2019

Our method

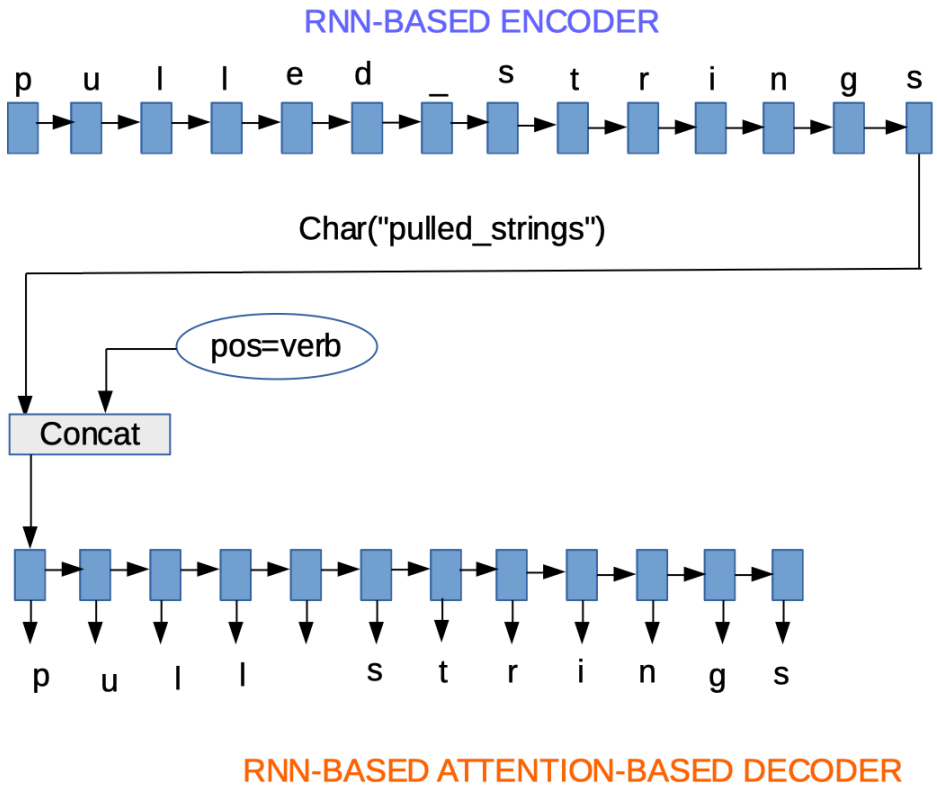
- Language-independent neural architecture
- Encoder-decoder approach $\rightarrow$ generation of MWE lemmas
- Exact-match evaluation
- Languages of experiment : French (pilot language, FR), Polish (PL), Italian (IT), Portuguese (BR, PT)

Datasets

- Many data sources and preprocessing scripts (cf. paper)
- Types : corpora (FR, PL) and dictionaries (all)
- Basic content : sets of (MWE form, base form)
- Potential other information associated with MWE form
- Addition of sets of (simple word form, simple word lemma) to capture single-word lemmatization information
- Size of training sets : FR (130k MWEs), PL (200k), IT (30k), PT (10k), BR (3k)

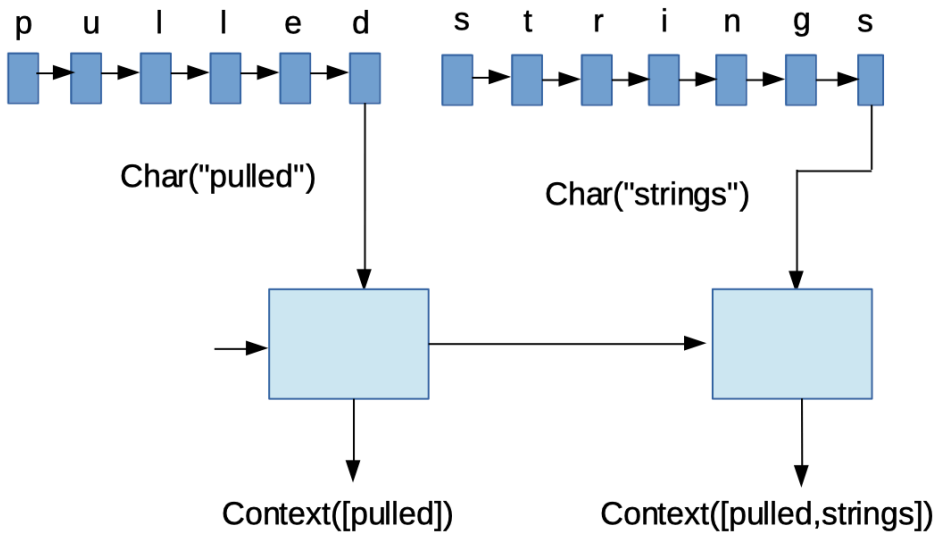
A brute-force encoder-decoder as a starter

(not in the paper)



- Low performances on French dev set
- But very good results for single-word lemmatization (97-99%)

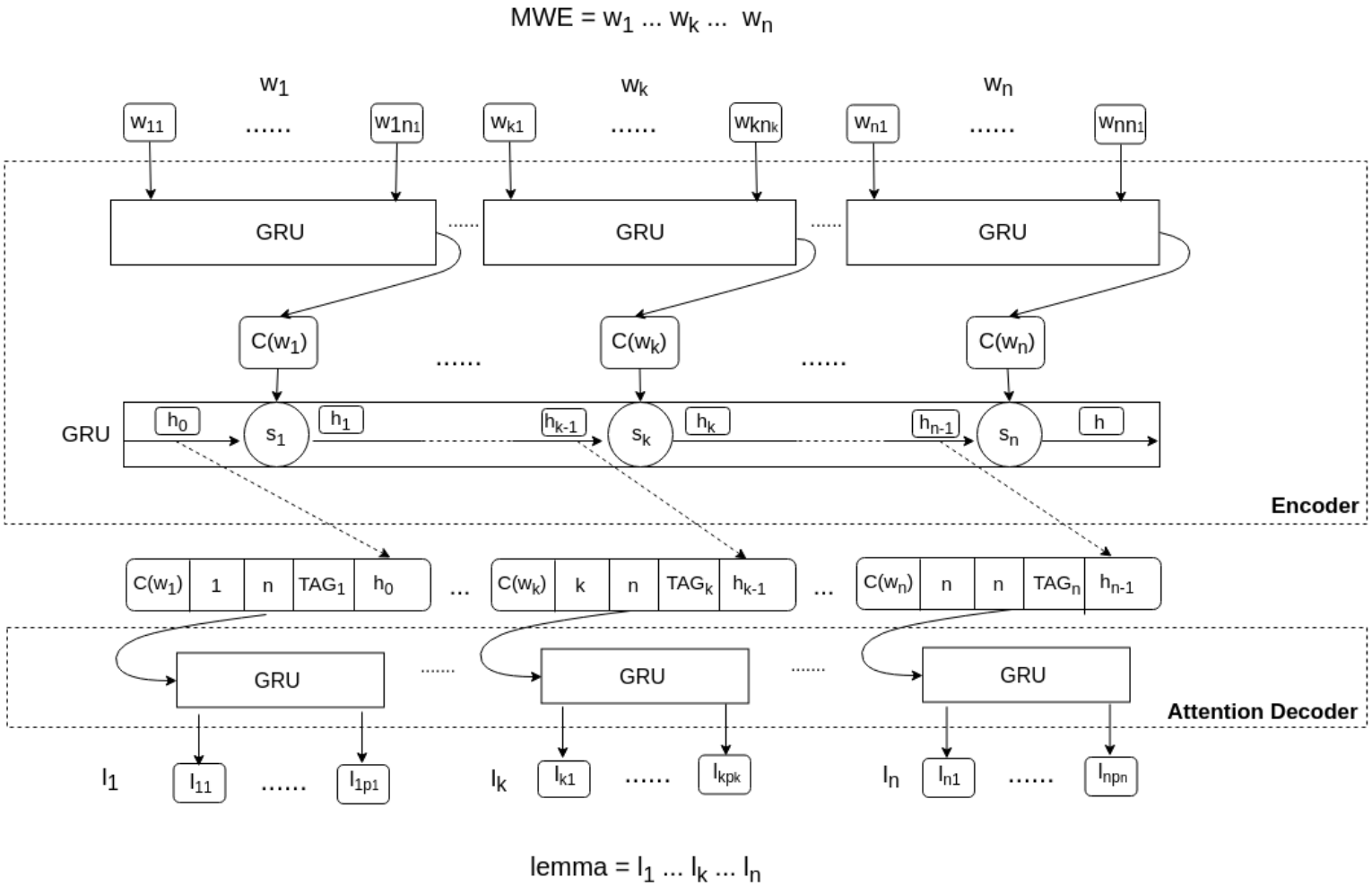
A two-level RNN-based encoder



Integration of tokenization in the MWE form

- character-level token encoder
- token sequence encoder

Global architecture



Results

	Test	
	all	unk.
FR corpus (non-verbal)	95.6	93.2
FR corpus (verbal)	75.2	75.2
FR dict (non-verbal)	87.5	88.4
PL corpus (nouns)	88.9	75.5
PL dict (mix)	58.6	59.0
IT (nouns)	91.7	91.7
PT (nominals)	88.2	88.4
BR (nominals)	81.6	81.6

- Not so bad results, but for Polish and for French verbal expressions
- ex. we could expect around 98% with a rule-based approach for Polish corpus (Marcinczuk, 2017)
- Good generalization over unknown multiwords

Comparing with a baseline

Baseline

UDPipe trained on our datasets using an IOB-like POS tagset

Results (on dev set)

	Our system	Baseline
FR corpus (non-verbal)	95.9	95.5
FR dict (non-verbal)	86.0	83.5
PL corpus (nouns)	88.9	70.1
PL dict (mix)	59.5	46.5

Our system consistently outperforms the baseline.

Ablation study

on French dev sets

	Dict	corpus (non-verb)
Complete system	86.0	95.9
- RNN on token sequence	75.6	88.1
- word POS tags	81.9	95.7
- position and length feats	83.6	95.8
- simple words in train set	78.3	88.9
Complete system + MWE gold tag	90.0	97.1
baseline UDPipe adaptation	83.5	95.5
baseline word-to-word	54.0	73.0

Conclusions

- Preliminary experiments on MWE lemmatization based on an encoder-decoder
- Not so bad results, but weak for Polish
- Future work : transformer-based approach, extraction of lemmatization rules + classification

THANKS FOR YOUR ATTENTION !
QUESTIONS ?

Results

	Dev (MWEs)		Test (MWEs)		Test (words)	
	all	unk.	all	unk.	all	unk.
FR ftb	95.9	91.5	95.6	93.2	98.0	96.8
FR shared task	73.1	73.1	75.2	75.2	82.7	82.6
FR dict	86.0	86.9	87.5	88.4	89.9	91.1
PL corpus	88.9	75.5	88.9	75.5	94.1	87.7
PL dict	59.5	59.5	58.6	59.0	76.8	76.8
IT	91.7	91.7	91.7	91.7	92.9	92.9
PT	89.7	89.7	88.2	88.4	95.1	95.1
BR	84.6	84.6	81.6	81.6	90.6	90.6

Other results

	French		Polish	
	Dict	Corp	Dict	Corp
(a) MWE lemma = MWE form	94.2 (*65.0*)	97.9 (*83.2*)	74.5 (*12.7*)	93.3 (*54.8*)
(b) MWE lemma = concat(lemmas)	95.8* (*55.8*)	99.4 (*70.4*)	67.4* (*28.5*)	90.9 (*43.1*)
Union of (a) and (b)	93.1 (*84.1*)	97.8 (*95.2*)	68.1 (*38.2*)	91.6 (*66.0*)
Intersection of (a) and (b)	99.1 (*35.2*)	100.0 (*62.5*)	85.5 (*3.0*)	93.4 (*31.9*)
Other MWE	82.5 (*15.9*)	85.7 (*4.8*)	57.3 (*61.8*)	83.2 (*34.0*)

TAB.: MWE-based accuracy on dev section according to MWE subclasses. * indicates that lemmas were predicted by UDPipe. Otherwise they are gold. Numbers between parentheses indicate the repartition of the MWE subclasses in the tested dataset (in percentage).