Lemmatisation d'expressions polylexicales par modèle neuronal

Mathieu Constant
Marine Schmitt

atilf ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

PARSEME.FR

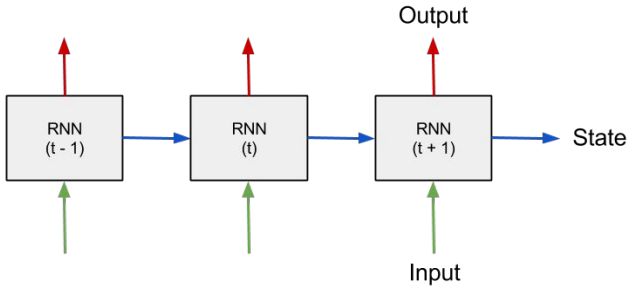
AGENCE NATIONALE DE LA RECHERCHE ANR

Introduction

- Lemmatisation d'expressions polylexicales de différents types (expressions nominales, adverbiales, verbales…) et dans plusieurs langues, principalement le français
- Enjeu de la tâche: extraire les propriétés morphologiques, lexicales et syntaxiques de l'expression pour sélectionner la bonne règle de lemmatisation (forme fixe, concaténation des lemmes, autre)
- Notre méthode: un réseau de neurones de type encodeur - décodeur qui génère, pour chaque mot de l'expression, le lemme correspondant

Les réseaux de neurones: vocabulaire

- RNN



$$q_{\text{mathematician}} = \left[\overbrace{2.3}^{\text{can run}}, \overbrace{9.4}^{\text{likes coffee}}, \overbrace{-5.5}^{\text{majored in Physics}}, \ldots \right]$$

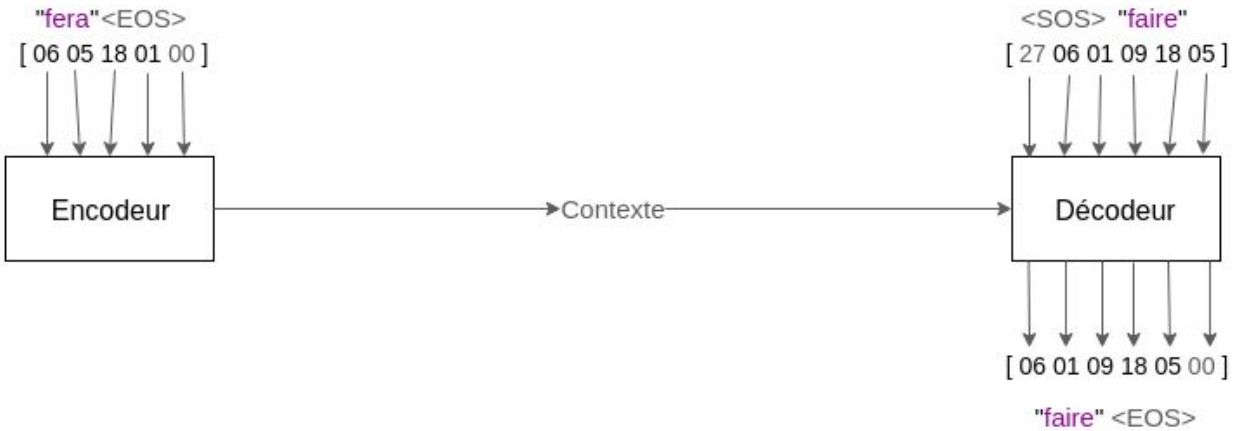
$$q_{\text{physicist}} = \left[\overbrace{2.5}^{\text{can run}}, \overbrace{9.1}^{\text{likes coffee}}, \overbrace{6.4}^{\text{majored in Physics}}, \ldots \right]$$

- Les embeddings : représentation d'une entité par un vecteur

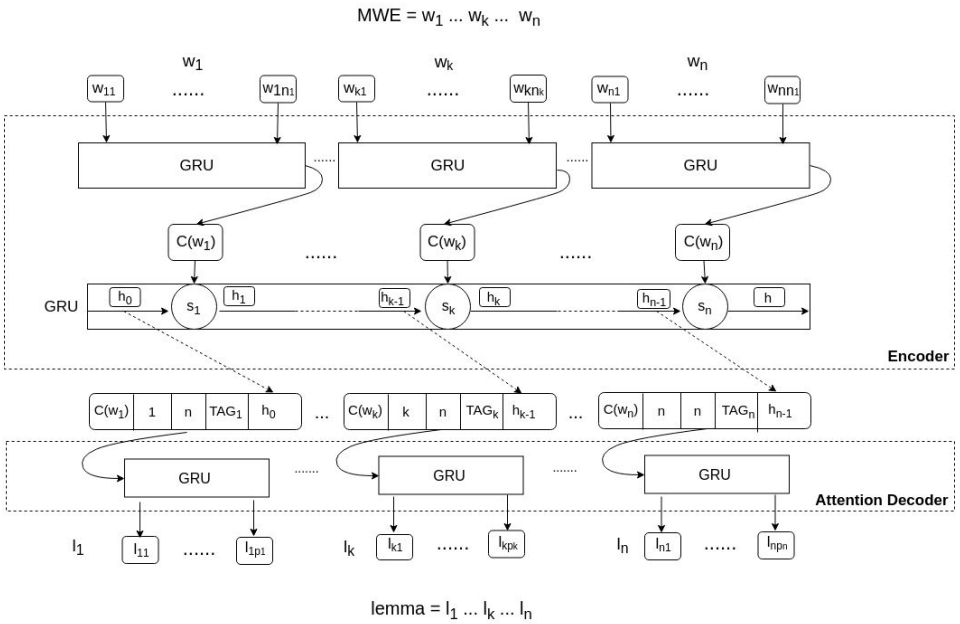
- GRU : - Pallier au problème du vanishing gradient + meilleur sur les longues séquences
  - Reset gate + Update gate -> déterminer l'information à garder et celle à oublier

Le modèle

- Lemmatisation d'un mot simple : encodeur/décodeur



- Lemmatisation d'une MWE : notre système



Les données

- 5 langues: français, polonais, italien, portugais, brésilien
- Corpus + Dictionnaires divisés en train/dev/test
- Pour le français, Corpus d'expressions verbales (Shared Task) appris séparément
- Ajout des mots simples pour le français et le polonais

Les expériences

- Paramétrage des expériences

  - ❏ Hyperparamètres: Taille des embeddings, taille du vecteur caché, taux d'apprentissage, dropout
  - ❏ Utilisation de UDPipe pour les POS tags prédits
  - ❏ Pour le polonais, ajout de l'information sur les cas (prédits par UDPipe)

- Les résultats

	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

- Bonne généralisation sur les mots inconnus
- Bonnes performances sur le français (expressions non verbales), l'italien, le portugais
- Bons scores sur le total des mots -> souvent un seul mot de l'expression mal lemmatisé
- Résultats décevants sur le polonais, sur les expressions verbales en français

Comparaison avec des baselines

Baselines:
- Adaptation de UDPipe : séquences de mots simples + MWEs avec POS tags + IOB tags
- Lemmatisation mot à mot avec UDPipe déjà entraîné

	Dict	FTB
Complete system	86.0	95.9
- GRU on word 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

- Le GRU sur la séquence des mots a une importance cruciale
- Le système a besoin de tous les composants pour atteindre son meilleur résultat
- Le système est amélioré par l'ajout de l'info sur le gold tag de l'expression
- Notre système montre de meilleures performances que les baselines

Résultats selon le type de MWE

	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*)

- Pour le français, le système est meilleur sur les expressions dont le lemme est la concaténation des lemmes
- Pour le polonais, il est meilleur sur les expressions fixes à la lemmatisation
- A noter, les excellents résultats sur l'intersection des deux catégories
- A noter, la grande proportion de MWEs n'appartenant à aucune des catégories pour le polonais (explication partielle des scores faibles)

Conclusion

- Bonnes performances de notre système sur le français
- Bonnes performances sur l'italien, le portugais et le brésilien malgré un manque de données
- Résultats décevants sur le polonais (langue morphologiquement riche et manque de données)
- Résultats particulièrement bons sur les expressions qui obéissent à l'une des deux règles principales mais le système apprend également sur les autres expressions
- Résultats meilleurs qu'une baseline solide (UDPipe)

Travail en cours

Réseau multilingue:

Mélange de plusieurs langues dans l'apprentissage pour pallier au manque de données sur certaines langues.
Résultats assez peu concluants jusqu'ici, sauf :

- dans le cas du portugais/brésilien (langues similaires)
- dans le cas du polonais pour 30k entrées ; polonais seul : ≈ 25%, polonais avec français et italien: ≈ 43%