À propos des devoirs1

felipe@iro.umontreal.ca

RALI
Dept. Informatique et Recherche Opérationnelle
Université de **Montréal**

V1.0

Last compiled: 13 novembre 2018

Plan

Participants

Plan

Participants

20 rapports remis

louis+antoine	Louis Clouatre, Antoine Colin	KN $<$ residual network
adel	Adel Nabli	KN + pos
kristof	Kristof Boucher Charbonneau	Kim2016
jean-francis	Jean-Francis Carignan	KN $\sim$ LSTM $<$ bi-LSTM
antoine	Antoine Gagnon	modèles n-gram
felix	Felix Martel	n-gram + char-based RNN
eric	Eric Girard	n-gram + RNN
antoine	Antoine Chehir	baseline
john+alexie	Johnathan Plante, Alexie Byrns	baseline
goergiy+amina	Georgiy Gegiya, Amina Madzhun	KN versus stupid + lemme
hugo	Hugo Lafortune-Brunet	KN
charles+orthlieb	Charles Ashby, Téo Orthlieb	baseline + RNN
julien+jeremy	Julien Allard, Jeremy Trudel	ngram + rule
yousra	Yousra Ben-Romdhane	baseline
philipp	Philipp Paquette	ngram + RNN
mouna+seb	Mouna salah, Sebastien Ehouan	LSTM + mySQL
tapopriya	Tapopriya Majumdar	word2vec
manish+shiven	Manish Kumar Jha, Shivendra Bhardwaj	ngram + RNN
alizee	Alizée Gagnon	ngram
christian	Christian Alaka	KN

Quelques points forts

- certains modèles état de l'art
  - louis+antoine Residual networks
  - kristof Kim et al. 2015
  - jean-francis bi-LSTM

kristof

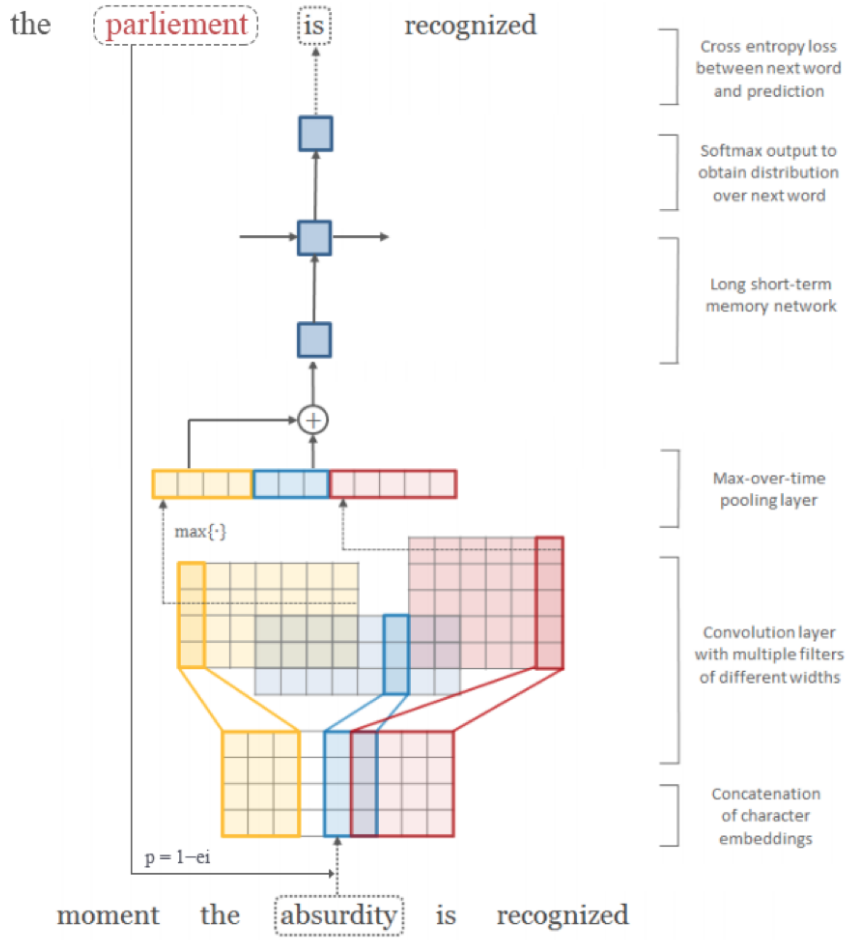

Figure 1. Architecture du réseau. En rouge, une mauvaise prédiction du réseau utilisé au temps actuel et en gris, une bonne prédiction. Inspiré de Kim et al. (2016)

Quelques points forts

- certains modèles état de l'art
  - louis+antoine Residual networks
  - kristof Kim et al. 2015
  - jean-francis bi-LSTM
- des versions modifiées de modèles connus
  - adel : modification de KN

Quelques points forts

- certains modèles état de l'art
  - louis+antoine Residual networks
  - kristof Kim et al. 2015
  - jean-francis bi-LSTM
- des versions modifiées de modèles connus
  - adel : modification de KN
- des rapports de très bonne qualité : jean-francis, kristof, adel, louis+antoine, felix, eric, /antoine

rali

Quelques points forts

- certains modèles état de l'art
  - louis+antoine Residual networks
  - kristof Kim et al. 2015
  - jean-francis bi-LSTM
- des versions modifiées de modèles connus
  - adel : modification de KN
- des rapports de très bonne qualité : jean-francis, kristof, adel, louis+antoine, felix, eric, /antoine
- des systèmes fonctionnels : john+alexie, goergiy+amina

rali

Quelques points forts

- certains modèles état de l'art
  - louis+antoine Residual networks
  - kristof Kim et al. 2015
  - jean-francis bi-LSTM
- des versions modifiées de modèles connus
  - adel : modification de KN
- des rapports de très bonne qualité : jean-francis, kristof, adel, louis+antoine, felix, eric, /antoine
- des systèmes fonctionnels : john+alexie, goergiy+amina
- un groupe (seulement) a effectué un **retour à la tâche** : yousra

rali

Quelques points forts

- certains modèles état de l'art
  - louis+antoine Residual networks
  - kristof Kim et al. 2015
  - jean-francis bi-LSTM
- des versions modifiées de modèles connus
  - adel : modification de KN
- des rapports de très bonne qualité : jean-francis, kristof, adel, louis+antoine, felix, eric, /antoine
- des systèmes fonctionnels : john+alexie, goergiy+amina
- un groupe (seulement) a effectué un **retour à la tâche** : yousra
- idée pour l'évaluation : utiliser la similarité des candidats aux mots à trouver : antoine

rali

Quelques points forts

- certains modèles état de l'art
  - louis+antoine Residual networks
  - kristof Kim et al. 2015
  - jean-francis bi-LSTM
- des versions modifiées de modèles connus
  - adel : modification de KN
- des rapports de très bonne qualité : jean-francis, kristof, adel, louis+antoine, felix, eric, /antoine
- des systèmes fonctionnels : john+alexie, goergiy+amina
- un groupe (seulement) a effectué un **retour à la tâche** : yousra
- idée pour l'évaluation : utiliser la similarité des candidats aux mots à trouver : antoine
- ajout de règles : julien+jeremy

rali

Baseline

I would also like to thank the shadow rapporteurs

1. faire glisser une fenêtre de taille 5 :

- - X - -	BOS	BOS	I	would	also
- - X	BOS	BOS	I		
X - -	I	would	also		
- - X - -	BOS	I	would	also	like
- - X	BOS	I	would		
X - -	would	also	like		
- - X - -	I	would	also	like	to
- - X	I	would	also		

2. ou 3 :

- X -	BOS	I	would
- X	BOS	I	
X -	I	would	
- X -	I	would	also
- X	I	would	

Baseline

- compter la fréquence de chaque contexte (`sort | uniq`)

43309	- - X	the	European	Union
43309	X - -	the	European	Union
48507	- - X	BOS	BOS	In
53039	- - X	Mr	President	,
53039	X - -	Mr	President	,
60238	- - X	BOS	BOS	It
73191	- - X	BOS	BOS	We
113028	- - X	BOS	BOS	I
129334	- - X	BOS	BOS	The
961118	X - -	.	EOS	EOS

$> 50M$ de règles pour l'anglais

- garder seulement le mot le plus fréquent de chaque contexte (mémoire safe)

  - - X BOS BOS The

Baseline

- ordonner les contextes de manière arbitraire (suite à de vagues tests sur `en-u5`)
  1. cc_cc
  2. c_c
  3. cc_
  4. c_
  5. _cc
  6. _c
  7. `The` (même en Finnois . . .)
- pas d'utilisation de la fréquence, pas de vote
- `baseline_rules-2+3`

Baseline

Those I have mentioned , and many others who accompanied them , could not <unk/> imagined the rapid <unk/> of their political project .

```
=> ccxcc: [could] [not] have [imagined] [the]
=> ccx: [could] [not] be
=> xcc: have [imagined] [the]
winner: CCXCC

=> ccx: [the] [rapid] reaction
=> xcc: because [of] [their]
winner: CCX
```

Métriques

good	% de phrases correctement traitées (incluant phrases avec un seul `unk`)
bad	% de phrases avec au moins une erreur
MPrec	pourcentage des `unk` correctement identifiés (moy. sur le corpus)
mPrec	moyenne des précisions par phrase

- influence de la casse
- prédictions numériques souple (`123` = `43` !)
- variantes n-best

rali

test : 18 remises

Adel-nabli	Kristof Boucher-Charbonneau
Alizée Gagnon	Manish Jha
Antoine Colin	Mouna Salah
Antoine Gagnon	Philip Paquette
Christian Alaka	Sebastien Ehouan
Georgiy Gegiya	Tapopriya Majumdar
Jeremy Trudel	Yousra Ben-Romdhane
Jonathan Plante	charles-ashby-teo-orthlieb
Julien Allard	Éric Girard

rali

toutes sortes de problèmes

/alakachr/test_tp1/alakachr-unk-europarl-v7.fi-en-u
<unk w="however"/> , in <unk w="view"/> of the shortage of <unk w="consumer"/> credit , there <unk w="is"/> a threat of energy and food crises . In order to maintain at least the current level of energy production , there will be a need by 2030 for worldwide investment of about USD 26 billion in reconstruction and the development of new oil and gas fields and also <unk w="promotes"/> the production and distribution of all types of energy . At <unk w="the"/> same time it will be necessary to integrate the flows of oil , gas and electricity so as to create an efficient and highly diversified system . This <unk w="policy"/> must help to. . .

- trop compliqué à aligner

toutes sortes de problèmes

- on garde les sorties ayant le bon nombre de lignes

```
wc -l allardj-trudejer3/allardj-trudejer-unk-healthca
968 allardj-trudejer3/allardj-trudejer-unk-healthcan
```

- conversion de format

```
In <unk w="the,this,addition"/> to <unk w="the,join,t
```

  - note : pas toujours facile `In <unk w="the,,",but"/> to`

```
s/"""/"@@@"/g
s?<unk/>?<unk w="@@@">?g
s?<UNK>?@@@?g
s?,,,?|@|?g
s?,,??g
s?\([^ ]\),\([^ ]\)?\1|\2?g
s?""?"?g
s?|"|??g
s?" "?"@"?g
```

In-domain : europarl-v7.fi-en-u5.en

straight	good	bad	MPrec	mPrec
julien+jeremy2	88	731	16.19	16.27
julien+jeremy1	98	716	17.18	17.41
julien+jeremy3	118	666	20.95	20.92
antoine2	120	664	21.28	21.27
antoine	120	664	21.28	21.27
eric	128	658	21.50	21.81
philipp	161	616	25.38	25.75
kristof-v3	177	605	26.70	27.72
adel	172	572	28.72	28.81
julien+jeremy4	180	577	28.12	28.64
baseline_rules-3	213	509	33.42	33.96
baseline_rules-2	225	504	34.74	34.83
goergiy+amina	230	513	34.46	34.91
louis+antoine	237	470	37.86	37.32
baseline_rules-2+3	260	460	38.68	39.26
goergiy+amina1	271	458	38.84	39.84
john+alexie	271	454	40.21	40.12

rali

john+alexie

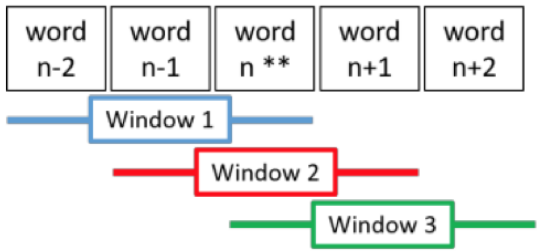

Figure 1. Example of windows created by the *sliding window* technique for trigram.

- combinaison de fenêtres de différentes tailles (2, 3 et 4)
- window1 et window2 prédisent souvent des mots différents
- inclure la ponctuation dans le vocabulaire semble aider (+1%)

goergiy+amina

- stupid backoff (implémentation personnelle) versus KN (KenLM)
- n-best prenant en compte la différence de probabilité avec le premier mot proposé

In-domain : europarl-v7.fi-en-u5.en

casse+number	good	bad	MPrec	mPrec
julien+jeremy2	105	698	18.65	18.87
julien+jeremy1	109	692	19.09	19.28
eric	128	658	21.50	21.81
julien+jeremy3	136	635	23.47	23.55
antoine2	139	636	23.69	23.80
antoine	139	636	23.69	23.80
philipp	161	616	25.44	25.79
kristof-v3	178	603	26.86	27.89
adel	178	560	29.76	29.77
julien+jeremy4	221	525	32.93	33.53
baseline_rules-3	213	508	33.64	34.08
baseline_rules-2	227	503	34.90	35.01
goergiy+amina	245	496	36.21	36.60
louis+antoine	239	468	38.07	37.53
baseline_rules-2+3	262	458	38.95	39.51
john+alexie	272	450	40.59	40.43
goergiy+amina1	326	408	44.80	45.60

rali

In-domain : europarl-v7.fi-en-u5.en

casse+number	good	bad	MPrec	mPrec	$MPrec_5$	$mPrec_5$
julien+jeremy2	105	698	18.65	18.87	29.76	30.23
julien+jeremy1	109	692	19.09	19.28	29.92	30.47
eric	128	658	21.50	21.81	32.28	32.86
julien+jeremy3	136	635	23.47	23.55		
antoine2	139	636	23.69	23.80	36.11	36.19
antoine	139	636	23.69	23.80	23.69	23.80
philipp	161	616	25.44	25.79	35.61	35.90
kristof-v3	178	603	26.86	27.89	40.32	41.47
adel	178	560	29.76	29.77	44.47	43.77
julien+jeremy4	221	525	32.93	33.53		
goergiy+amina2	245	496	36.21	36.60	43.98	44.48
louis+antoine	239	468	38.07	37.53		
baseline_rules-2+3	262	458	38.95	39.51		
john+alexie	272	450	40.59	40.43	46.94	46.70
goergiy+amina1	326	408	44.80	45.60	46.39	47.45

In-domain : europarl-v7.fi-en-u50.en

casse+number	good	bad	$MPrec_1$	$mPrec_1$
eric	0	404	7.46	7.96
antoine2	0	260	13.53	13.66
antoine	0	260	13.53	13.66
kristof-v3	1	221	13.70	14.02
goergiy+amina2	1	231	14.18	14.92
philipp	1	238	14.07	14.33
baseline_rules-3	1	149	17.43	17.99
baseline_rules-2	1	116	20.93	21.29
john+alexie	3	113	21.82	22.76
louis+antoine	2	133	22.66	22.21
baseline_rules-2+3	2	96	23.09	23.64

Out-domain : hans-20.en

casse+number	good	bad	$MPrec_1$	
eric	19	525	12.35	
antoine2	17	476	15.81	
antoine	17	476	15.81	
philipp	21	466	16.18	
baseline_rules-3	27	379	20.29	
baseline_rules-2	30	342	21.58	
baseline_rules-2+3	35	311	24.62	
john+alexie	40	310	24.78	
louis+antoine	31	307	25.01	
goergiy+amina1	48	311	25.26	
goergiy+amina1	80	198	33.43	europarl-u20

Out-domain : healthcan-20.en

casse+number	good	bad	$MPrec_1$	
eric	6	736	8.55	
antoine2	15	663	12.34	
antoine	15	663	12.34	
baseline_rules	13	630	13.68	
kristof-v3	16	602	14.40	
louis+antoine	17	565	17.53	
baseline_rules-2	23	578	16.88	
baseline_rules-2+3	26	567	17.81	
goergiy+amina2	25	561	17.89	
john+alexie	28	559	17.78	
goergiy+amina1	28	556	18.42	
goergiy+amina1	48	311	25.26	hans-20