



Présentation finale :

Application des réseaux de neurones au traitement des langues naturelles

Elèves:

Hamza Ed-dbiri

Said Khaboud

Hugo Mailfait

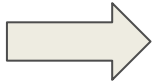
Salaheddine Mesdar

Tuteur :

Alexandre Saidi

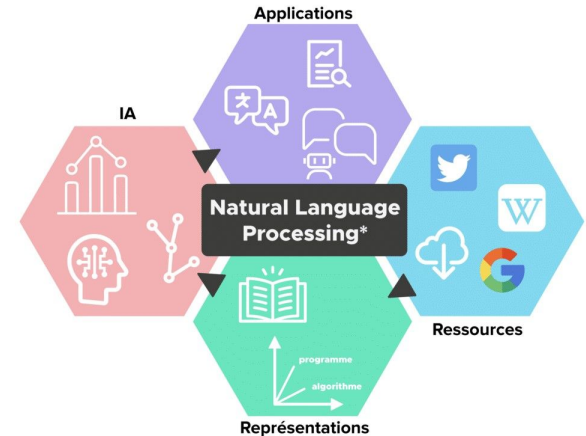
Appropriation du sujet

Qu'est-ce que le *Natural Language Processing* (NLP) ?



Il s'agit des techniques visant à modéliser et reproduire la capacité humaine à produire et à comprendre des énoncés linguistiques.

- Intérêt = Obtenir de nouvelles informations grâce à l'explosion des données disponibles en ligne.
- Difficulté principale = Ambiguïté / Implicité
- Applications principales actuelles = Résumés automatiques, reconnaissance vocale, Q&A, traduction, classification de textes ...



Objectifs du projet

- ❑ Se familiariser avec les techniques de l'IA en Text Mining et NLP (recherches personnelles, rapports d'anciens élèves, bibliographie proposée par notre tuteur ...).
- ❑ Réaliser un **chatbot** sous la forme de questions/réponses à partir d'une ou plusieurs technique issues de l'état de l'art
- ❑ Livrables: Rapport final + Prototypes

Sommaire

- 1) Choix du corpus
- 2) Le *preprocessing* en NLP
- 3) Présentation des modèles *retrieve*
- 4) Présentation du modèle *generative*
- 5) Evaluation des modèles et perspectives

1. Choix du corpus

Initialement → **Règlement de la scolarité** et fiches sur le cursus à l'ECL



Dataset limité



17 décembre 2003 / 3h 21min / Fantastique, Aventure

De [Peter Jackson](#)

Avec Sean Astin, Elijah Wood, Viggo Mortensen

Titre original The Lord of the Rings: The Return of the King



PRESSE



18 critiques

SPECTATEURS



112058 notes dont 3578 critiques

MES AMIS



2. Etape du preprocessing

I feel so LUCKY to have found this (used) phone.

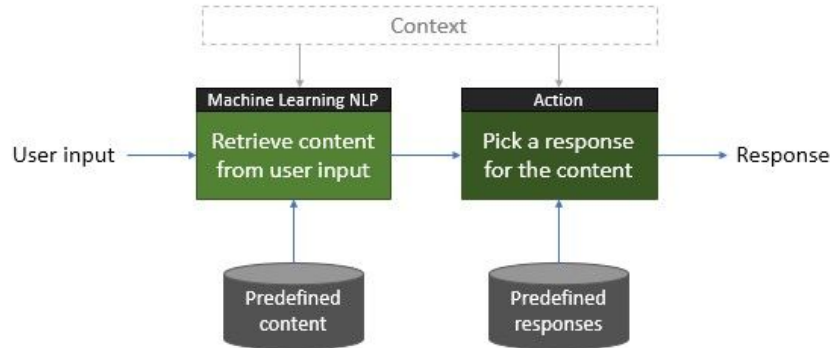
Eliminer la ponctuation	I feel so LUCKY to have found this used phone
Convertir en minuscules	i feel so lucky to have found this used phone
Tokenization	[i, feel, so, lucky, to, have, found, this, used, phone]
Eliminer les StopWords	[feel, lucky, have, found, used, phone]
Lemmatization	[feel, lucky, have, find, use, phone]

Construction des modèles

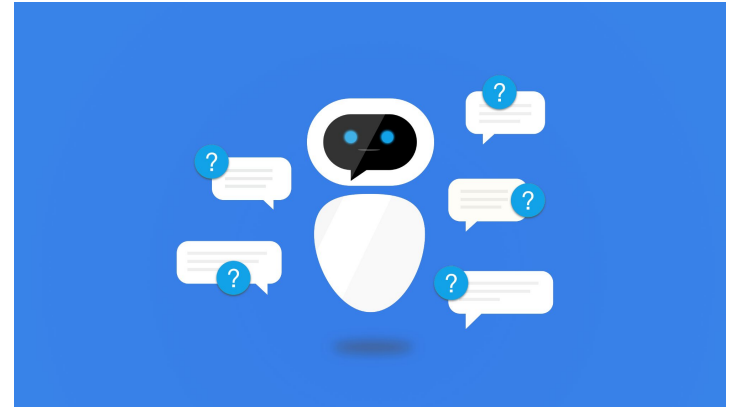
Retrieve

VS

Generative



Meilleure réponse possible
parmi une base de données de
réponses pré-définies



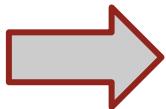
Réponse créée à partir d'un
réseau de neurone Seq2Seq

3. Les modèles *retrieve* (exemple 1)

Le principe TF-IDF

Term Frequency

Fréquence d'apparition du
mot dans le document



	mot	année	rêve	type	enfant
DOC1	0	0.21	0.08	0	0
DOC2	0.08	0	0	0	0
DOC3	0.13	0	0	0	0.13

Inverse Document Frequency

Diminue les fréquences des
mots présents dans de
nombreux documents

$$idf_1 = \log \left(\frac{|D|}{|\{d_j : t_1 \in d_j\}|} \right)$$

3. Les modèles *retrieve* (exemple 1)

Corpus de questions prédéfinies

```
{
  "intent": "synopsis",
  "examples": [
    "décris moi le film",
    "decris le film",
    "description",
    "détails du film",
    "details",
    "dis m'en plus sur ce film",
    "description du film",
    "de quoi parle le film ?",
    "quel est le synopsis ?",
    "histoire racontée",
    "de quoi parle le film ?",
    "résumé du film",
    "resume"]
},
{
  "intent": "box_office",
  "examples": [
    "nombre d'entrées",
    "entrées",
    "entrees",
    "box office",
    "revenus du film",
    "combien a rapporté le film ?",
    "combien d'entrées au box-office ?",
    "combien de personnes ont vu ce film au cinéma ?",
    "le film est-il populaire ?"]
},
```

- Ajout de la nouvelle question au corpus
- Module scikit-learn *TfidfVectorizer* pour le calcul de la matrice TF-IDF



*Cos-similarité
de la nouvelle
question avec
la i-ème
question*



**On récupère la
question avec la
similarité la plus
proche**

3. Les modèles *retrieve* (exemple 1)

Détection du titre

Bonjour, je suis Alex, votre chatbot AlloCine.
Je peux répondre à toutes vos questions sur de nombreux films : le réalisateur, le genre, la durée, la note des spectateurs, le synopsis etc ...

Tout d'abord, pourriez-vous rentrer le film sur lequel vous souhaitez des informations :
le silence des agneaux

Le film choisi est-il l'un des trois ci-dessous ?
['Le Silence des agneaux', 'Le Labyrinthe du silence', 'Les Jeux des nuages et de la pluie']

Si oui, indiquez son numéro dans la liste, si non écrivez "erreur" : 1

Questions sur le film

Quelle est votre question ? : réalisateur
Jonathan Demme

Quelle est votre question ? : qui sont les acteurs principaux ?
Anthony Hopkins, Jodie Foster, Scott Glenn

Quelle est votre question ? : quel score au box-office ?
3 119 085 entrées

Quelle est votre question ? : QUEL BUDGET
\$19.000.000 dlls

Quelle est votre question ? : au revoir
Au revoir et à bientôt j'espère !

3. Les modèles *retrieve* (exemple 2)

Bag-of-words (BOW)

Méthode de vectorisation de texte très utilisée en text mining

Document	the	cat	sat	in	hat	with
<i>the cat sat</i>	1	1	1	0	0	0
<i>the cat sat in the hat</i>	2	1	1	1	1	0
<i>the cat with the hat</i>	2	1	0	0	1	1

3. Les modèles *retrieve* (exemple 2)

Création du corpus

```
{ "tag": "Hopeful Notes title",  
  "patterns": ["title of Hopeful Notes", "What is the title of Hopeful Notes", "title of Hopeful Notes?"],  
  "responses": ["Hopeful Notes"],  
  "context": [""],  
  { "tag": "Hopeful Notes year",  
    "patterns": ["year of Hopeful Notes", "What is the year of Hopeful Notes", "year of Hopeful Notes?"],  
    "responses": ["2010"],  
    "context": [""],  
    { "tag": "Hopeful Notes date",  
      "patterns": ["date of Hopeful Notes", "What is the date of Hopeful Notes", "date of Hopeful Notes?"],  
      "responses": ["2010-12-15"],  
      "context": [""],  
      { "tag": "Hopeful Notes genre",  
        "patterns": ["genre of Hopeful Notes", "What is the genre of Hopeful Notes", "genre of Hopeful Notes?"],  
        "responses": ["Drama"],  
        "context": [""],  
        { "tag": "Hopeful Notes duration",  
          "patterns": ["duration of Hopeful Notes", "What is the duration of Hopeful Notes",  
            "duration of Hopeful Notes?"],  
          "responses": ["94"],  
          "context": [""],  
          { "tag": "Hopeful Notes country",  
            "patterns": ["country of Hopeful Notes", "What is the country of Hopeful Notes", "country of Hopeful Notes?"],  
            "responses": ["USA"],  
            "context": [""],  
            }
```

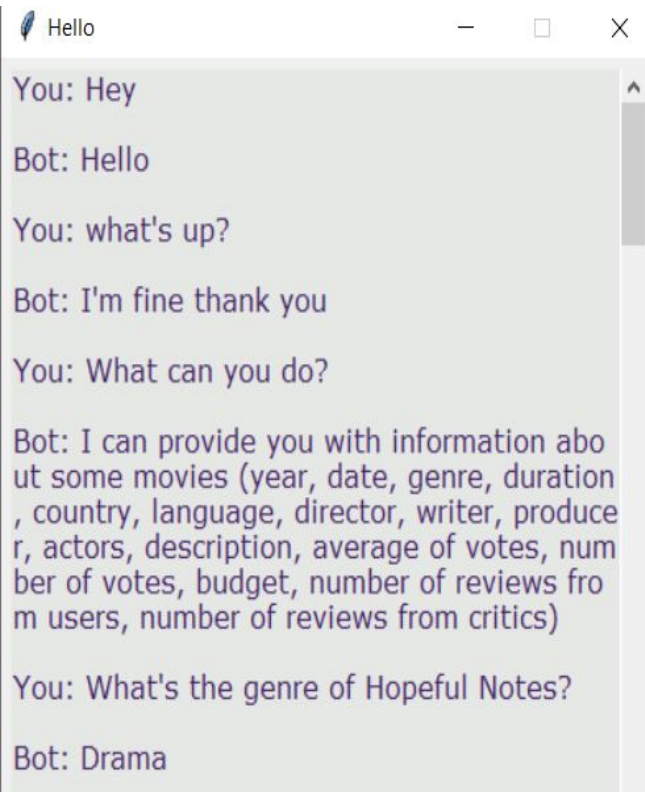
3. Les modèles *retrieve* (exemple 2)

Développement d'un modèle et génération des réponses

- Utilisation d'un réseau de neurones simple (un perceptron multicouches)
- Utilisation de la descente de gradient stochastique
- Création des classes pour les intentions
- Génération d'une réponse à l'utilisateur

3. Les modèles *retrieve* (exemple 2)

Interactions avec le chatbot



Chatbot window 1 (Hello):

You: Hey

Bot: Hello

You: what's up?

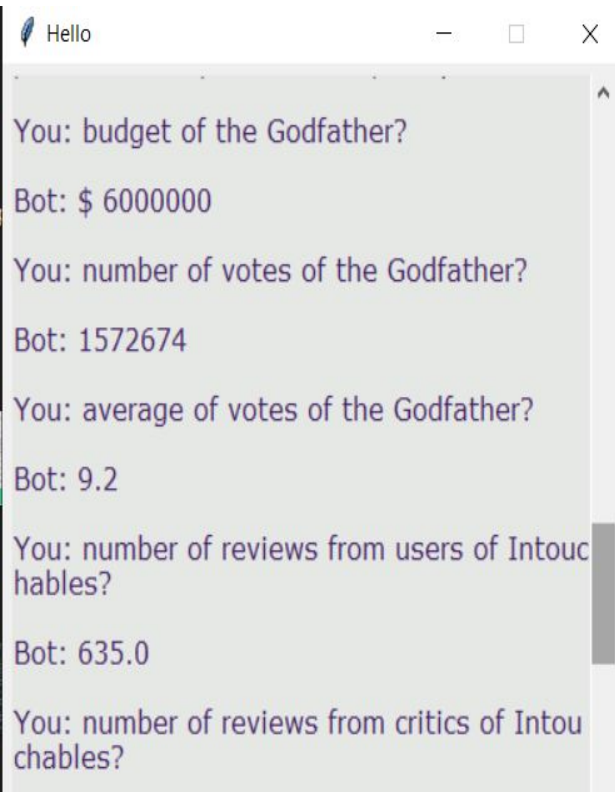
Bot: I'm fine thank you

You: What can you do?

Bot: I can provide you with information about some movies (year, date, genre, duration, country, language, director, writer, producer, actors, description, average of votes, number of votes, budget, number of reviews from users, number of reviews from critics)

You: What's the genre of Hopeful Notes?

Bot: Drama



Chatbot window 2 (Hello):

You: budget of the Godfather?

Bot: \$ 6000000

You: number of votes of the Godfather?

Bot: 1572674

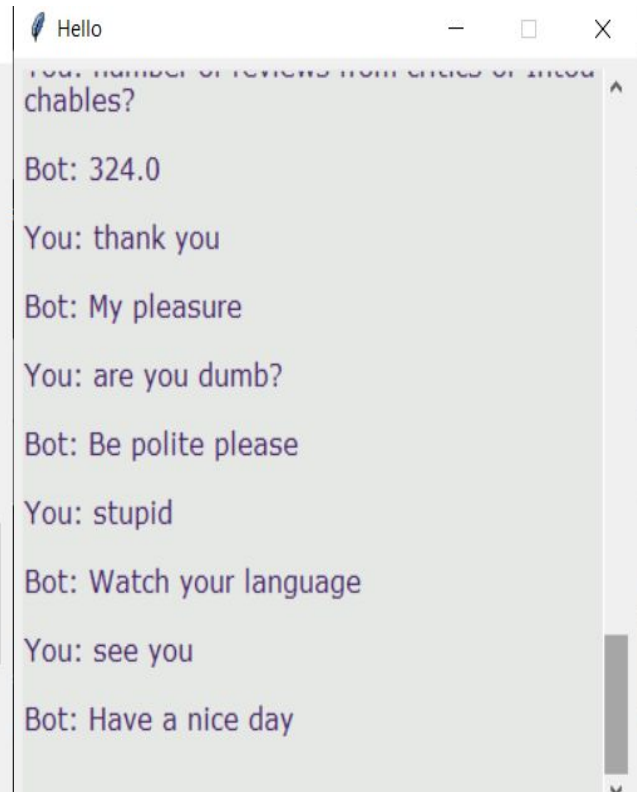
You: average of votes of the Godfather?

Bot: 9.2

You: number of reviews from users of Intouchables?

Bot: 635.0

You: number of reviews from critics of Intouchables?



Chatbot window 3 (Hello):

You: number of reviews from critics of Intouchables?

Bot: 324.0

You: thank you

Bot: My pleasure

You: are you dumb?

Bot: Be polite please

You: stupid

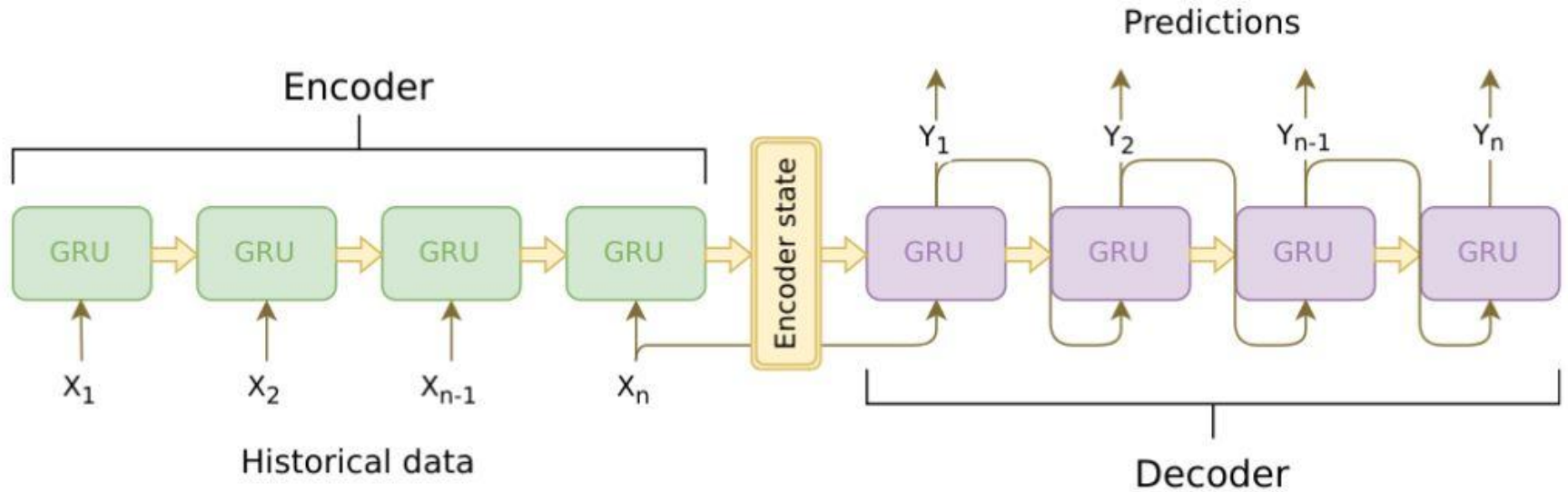
Bot: Watch your language

You: see you

Bot: Have a nice day

4. Le modèle *generative*

Utilisation d'un modèle Seq2Seq

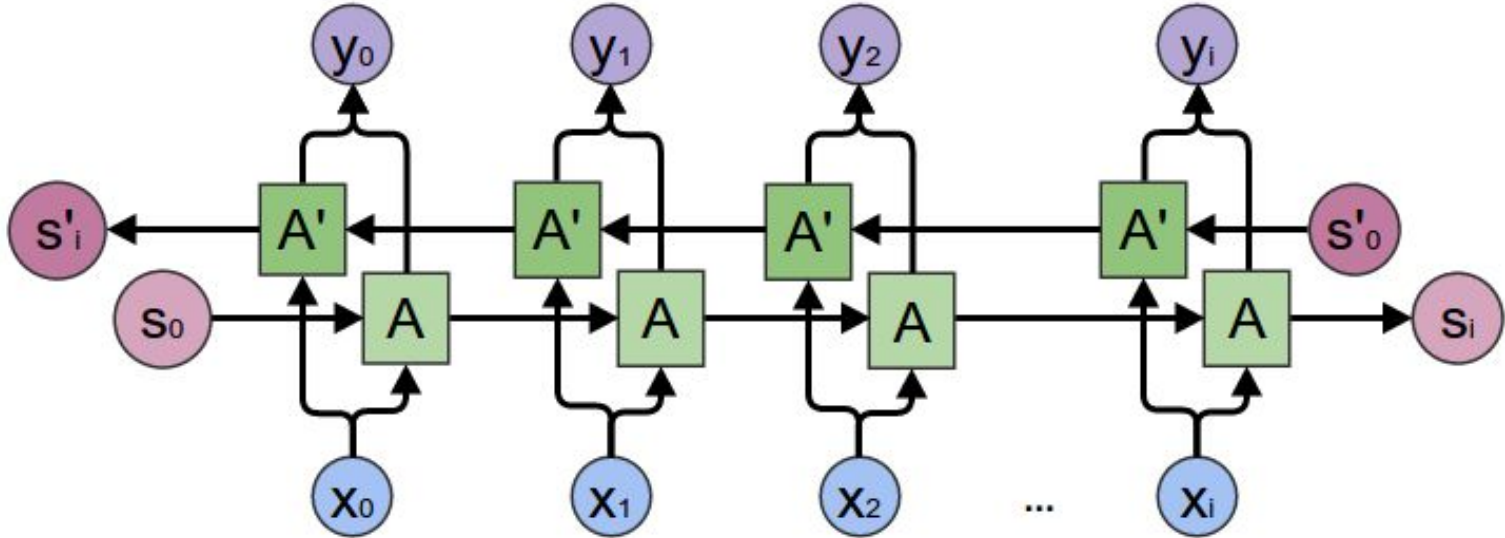


4. Le modèle *generative*

Encoder



GRU bidirectionnel

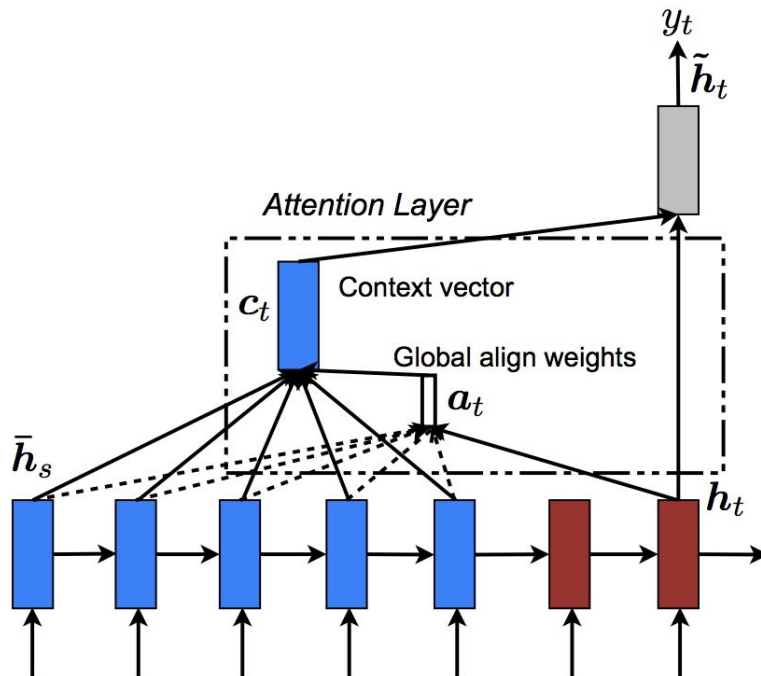


4. Le modèle *generative*

Decoder



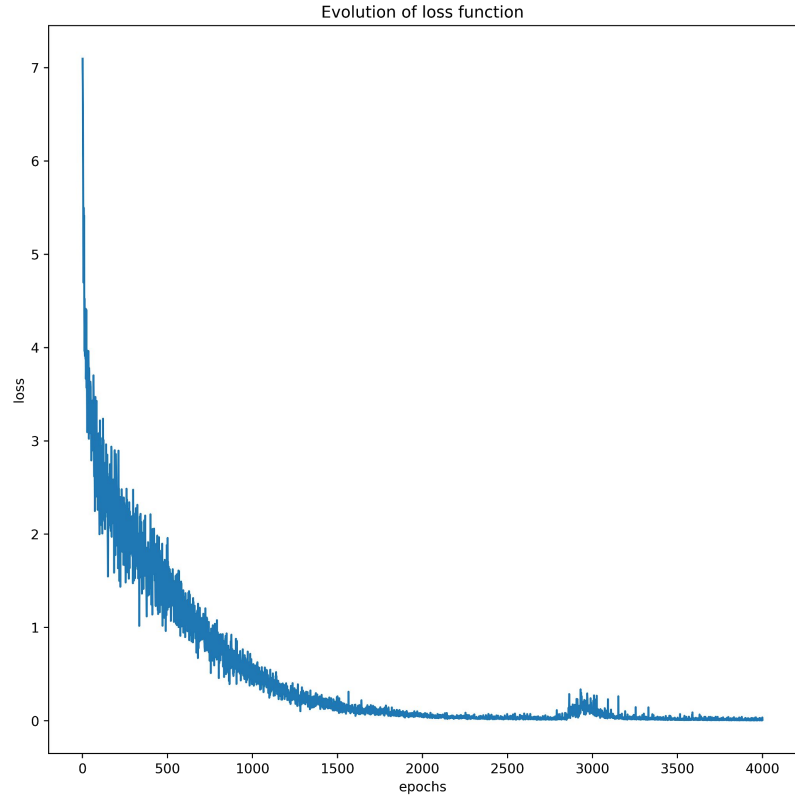
GRU avec attention



4. Le modèle *generative*

Phase d'entraînement

Loss function avec application
du teacher forcing et le
gradient clipping



4. Le modèle *generative*

Interrogation



Greedy search decoder

```
evaluateInput(encoder, decoder, searcher, voc)
```

```
> hello
Bot: hello
> how are you
Bot: good and you ?
> language of amercian beauty
Error: Encountered unknown word.
> language of american beauty
Bot: english
> country of american beauty
Bot: usa
> budget of the godfather
Bot: $ 6000000
> number of votes of the godfather
Bot: 1572674 deric
> see you
Bot: bye ! come back again soon .
> quit
```

```
> hello
Bot: hello
> are you dumb ?
Bot: watch your language
> stupid
Bot: be polite please
> thank you
Bot: bye ! come back again soon .
> quit
```

5. Evaluation des modèles

Limitations et points d'amélioration du modèle *generative* :

- Génération de longues séquences compliquée.
- Mots non rencontrés
 - "Error: Encountered Unknown Vector"
 - Taille du corpus de questions/réponses à augmenter
- Amélioration de la précision avec des embeddings pré-entraînés.

5. Evaluation des modèles

Modèle TF-IDF

- + Simple, efficace et rapide
- + Prise en main facile
- Aspect robotique inhérent au retrieve
- Ultra-dépendance du modèle aux questions pré-établies
- Quelques imprécisions dans les réponses

Ambiguïté Note-Critique

```
Quelle est votre question ? : critique  
50655 notes dont 2844 critiques
```

Mauvaise réponse

```
Quelle est votre question ? : quel sujet ?  
4,3
```

5. Evaluation des modèles

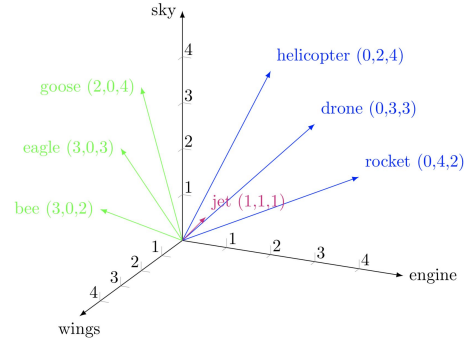
Modèle Bag-of-words

- + Simple, efficace et rapide
- + Prise en main facile
- + Interaction vive avec l'utilisateur
- + Précision de 69.98%
- Sens sémantique
- Taille du vecteur
- Absence de réponse en cas d'incompréhension de la question de l'utilisateur

5. Perspectives

- Adapter le modèle Word2Vec à des phrases complètes

Word Embedding



- Système de recommandation de films



**Merci de votre
attention**