Evaluation des “performances” des systèmes de TAL

Benoit Favre, Frédéric Bechet

Aix-Marseille Université - LIS

Master M1

Evaluation

- L'évaluation est au coeur des méthodes empiriques basées sur l'apprentissage automatique
- Un *modèle* n'existe pas pour lui-même, pour ce qu'il est capable d'expliquer sur le *monde*, mais uniquement par rapport à sa **performance**
- La **performance** permet de mesurer à quelle point un modèle *simule* un phénomène *naturel*
- La **performance** est aussi au coeur des processus d'apprentissage $\rightarrow$ tout le processus d'apprentissage est basé sur la minimisation de l'erreur d'apprentissage

Mais que signifie performance pour des tâches linguistiques ? et comment est elle estimée ?

Performance et TAL

- Paradigme d'évaluation : **l'humain est la référence**
  - tout processus de TAL pour une tâche T a pour objet de reproduire le comportement d'un humain réalisant cette tâche T
  - la **performance** correspond à l'écart de comportement entre le processus automatique et l'humain
  - exemple : classification morphosyntaxique (POS) des mots d'un texte
    - un *expert* humain effectue cet étiquetage *à la main*
    - un système automatique produit un étiquetage qui est comparé à l'étiquetage manuel
    - performance = taux de bonne classification

Performance et TAL

- Problèmes du paradigme : **l'humain est la référence**
  1. quel humain ?
  2. quel référence ?
- Tous les humains n'ont pas le même degré d'expertise linguistique
  - problème de la *fiabilité* des annotations humaines
  - crucial à l'heure des annotations par *crowdsourcing* (ex : *Amazon Mechanical Turk*)
- Pour de nombreuses tâches il n'y a pas de référence unique
  - langue naturelle *ambigüe* et *implicite*
  - tâches de *génération de texte* $\rightarrow$ traduction, résumé automatique
  - problème pour les tâches interactives $\rightarrow$ dialogue humain-machine

Performance et TAL

- Quel niveau d'analyse ?
  1. évaluations *extrinsèques* $\rightarrow$ liées à la tâche finale
     - ex : succès d'une réservation pour un serveur vocal de réservation d'hotels
     - tâche = traitement de signal + transcription + compréhension + gestion dialogue + génération + synthèse vocale
     - avantage : évaluation directement lié à l'*utilité* d'un modèle
     - inconvénient : difficile à utiliser comme *fonction objective* dans la phase d'apprentissage de chaque sous-tâche !
  2. évaluations *intrinsèques* $\rightarrow$ chaque sous-tâche est évaluée indépendemment
     ex : transcription / compréhension / compréhension en contexte / prédiction de l'action suivante / génération / ...
     - avantage : utilisation de la même métrique pour l'évaluation et l'apprentissage
     - inconvénients : évaluation parfois *artificielle*, éloignée de l'objectif final

Référence

La **référence** est la classe/étiquette déterminée par un ou plusieurs humains pour un exemple donné ($y^\star$).
L'évaluation automatique :

- Comparaison des sorties du système à la "verité"
- Permet une boucle "amélioration $\rightarrow$ évaluation $\rightarrow$ amélioration $\rightarrow$ évaluation ..."
- Requiert un corpus **annoté**

L'évaluation manuelle :

- Lorsqu'il n'existe pas de référence (ex : résumé, traduction)
- Lorsque le processus est trop complexe
- Lorsque l'annotation est trop chère
- Lorsque la mesure automatique n'est pas fiable

Mesures classiques

- Taux d'erreur, de succès

$$err = \frac{nb(erreurs)}{nb(reference)} \qquad corr = \frac{nb(juste)}{nb(reference)} \qquad err = 1 - corr$$

Mesures classiques

	Ref	Non-Ref
Hyp	Vrai positifs	Faux positifs
Non-Hyp	Vrai négatifs	Faux négatifs

Mesures classiques

- Rappel-précision

$$rappel = \frac{nb(juste)}{nb(reference)}$$

$$precision = \frac{nb(juste)}{nb(hypothese)}$$

- F-score

$$\text{F-score} = \frac{2 \times precision \times rappel}{precision + rappel}$$

Pourquoi le taux d'erreur peut-il être trompeur ?

- Le taux d'erreur est pertinent lorsque *tous* les éléments d'un corpus reçoivent une étiquette
- Dans de nombreuses tâches, il faut d'abord détecter les éléments à étiqueter, puis ensuite trouver l'étiquette
  - le taux d'éléments sans étiquettes peut avoir une influence trop grande sur le taux d'erreurs
  - exemple :

```
0       investiture     nc      O
1       aujourd'hui     adv     O
2       à               prep    O
3       Bamako          np      B-geoloc
4       ,               ponctw  O
5       Mali            np      B-geoloc
6       ,               ponctw  O
7       du              prep    O
8       président       nc      O
9       Touré           np      B-person
```

un système qui prédit *toujours* l'étiquette *O* est à correct à 70%

exemple : les entités nommées

```
EXAMPLE ERROR RATE (argmax): 0.060568 (954/15751)
  B-geoloc:  r=0.46  p=0.86  f=0.60 (tp=197,   fp=32,  pp=426)
  B-org:     r=0.42  p=0.79  f=0.55 (tp=105,   fp=27,  pp=246)
  B-person:  r=0.49  p=0.91  f=0.64 (tp=134,   fp=12,  pp=270)
  B-product: r=0.05  p=0.50  f=0.10 (tp=1,     fp=1,   pp=17)
  O:         r=0.99 p=0.94   f=0.97 (tp=14232, fp=816, pp=14283)
```

- taux d'erreur global : **6%**
- taux de classification correcte : **94%**
- résultats sur les entités nommées
  - Macro-F1 (moyenne des F1) : **0,47**
    - ★ calcul : $(0,60 + 0,55 + 0,64 + 0,10)/4$
  - Micro-F1 (moyenne pondérée par le nombre d'exemples) : **0.58**
    - ★ prec=$(197 + 105 + 134 + 1)/(197 + 105 + 134 + 1 + 32 + 27 + 12 + 1) = 0,85$
    - ★ rapp=$(197 + 105 + 134 + 1)/(426 + 246 + 270 + 17) = 0,45$

plusieurs sources d'erreurs

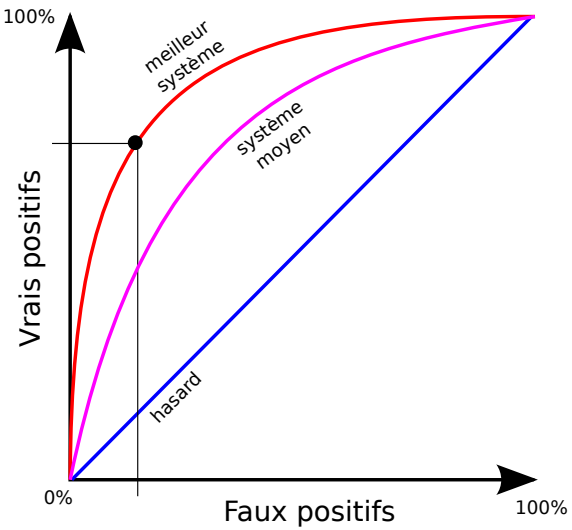
- il peut y avoir plusieurs sources d'erreurs qui se cumulent
  - exemple : erreurs de segmentation / erreur d'étiquetage

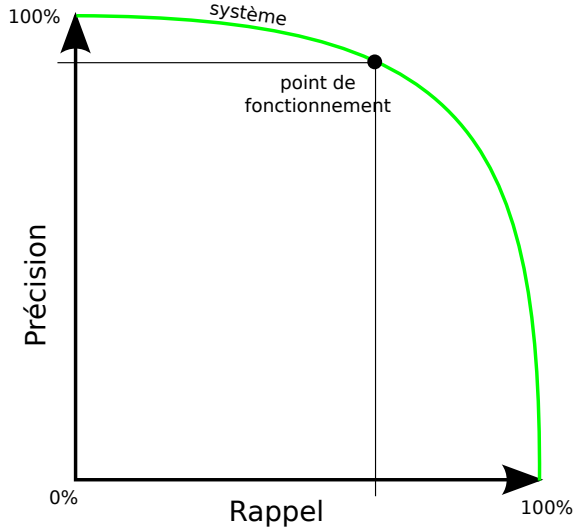
indice	mot	ref	syst1	syst2
0	investiture	O	O	O
1	aujourd'hui	O	O	O
2	à	O	O	B-geoloc
3	Bamako	B-geoloc	B-geoloc	I-geoloc
4	,	O	O	O
5	Mali	B-geoloc	B-org	B-geoloc
6	,	O	O	O
7	du	O	O	O
8	président	O	O	B-person
9	Touré	B-person	B-org	I-person

- Si l'on prend en compte l'évaluation globale, les systèmes 1 et 2 sont équivalents :
  - 1 succès, 2 erreurs chacun
- Sont-ils vraiment équivalents ?
  - pondérations par *type* d'erreurs

Courbes

- ROC (Receiver Operating Characteristic), AUR (area under the ROC curve)
  - Faux positifs / Vrai positifs (=Rappel)





- Point de fonctionnement : choix d'une configuration adaptée au déploiement

Évaluation manuelle

- Notes graduelles
  - "Les sorties du système correspondent-elles à vos attentes ?" : 1-2-3-4-5
  - Limites : tout le monde n'interprète pas l'échelle de la même façon.
- Comparaisons
  - "Les sorties de A sont-elles meilleures que celles de B ?" : oui / non
  - Limites : $(A > B)$ et $(B > C)$ n'implique pas forcément $(A > C)$
- Évaluation implicite
  - Succès sur une tâche pour laquelle l'utilisateur tire parti des sorties du systèmes
  - Exemple : rechercher une information sur un site web en Chinois avec/sans traduction automatique.
  - Limites : évaluation globale de tous les composants impliqués

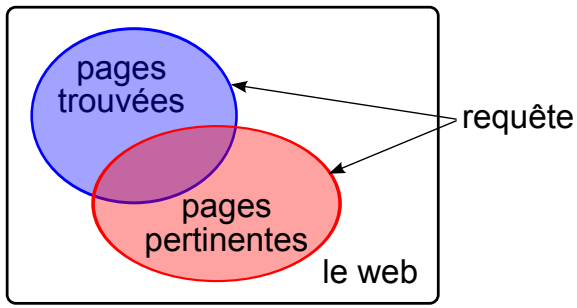
Recherche d'information

Comment évaluer les résultats d'un moteur de recherche ?

- Définir des requêtes, puis étiqueter **tous** les documents (pertinent / non-pertinent)

Mesures :

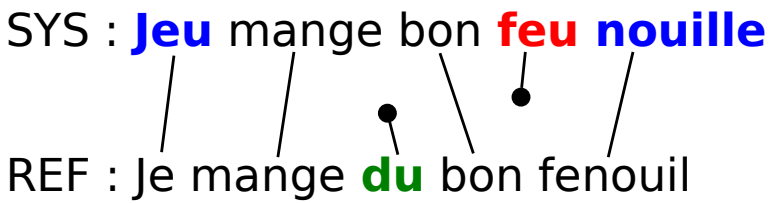
- Rappel-précision
- Précision moyenne
- Précision à 10 documents



Transcription de parole

- Entrées = audio, sorties = séquence de mots
- Taux d'erreur mots (Word Error Rate)

$$\text{wer} = \frac{n(\text{ins})) + n(\text{sub}) + n(\text{del})}{n(ref)}$$



- Alignement déterminé par programmation dynamique

Evaluation de tâches de génération de textes

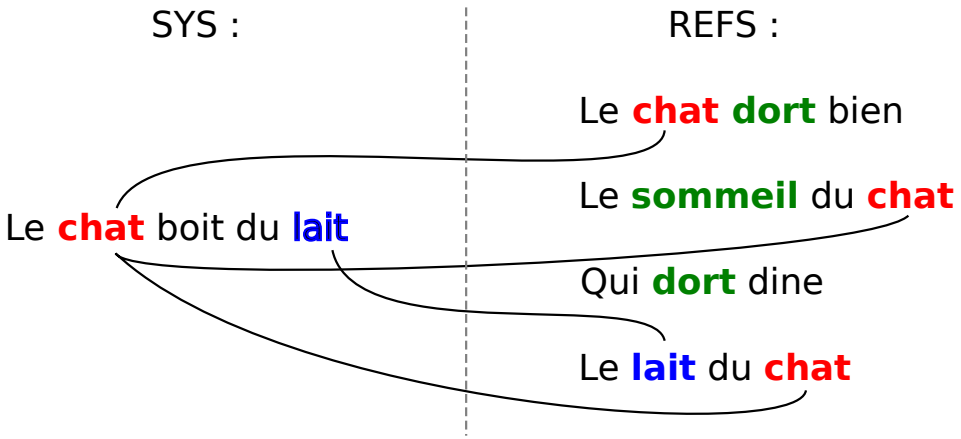
- Génération de texte : traduction, résumé automatique
- Problèmes
  - il n'y a pas de références uniques ! !
  - comment mesurer automatiquement la *qualité* d'un texte ?
- Solutions
  - utiliser des références multiples (le plus possible ! !)
  - utiliser des métriques *corrélées* (plus ou moins) avec l'évaluation recherchée

Résumé automatique

- Entrées = document(s) et longueur voulue, sorties = texte < longueur
- Manuelle :
  - Fond : 1-5 (événements, dates, acteurs principaux)
  - Forme : 1-5 (grammaire, organisation, clarté...)
- Automatique : ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

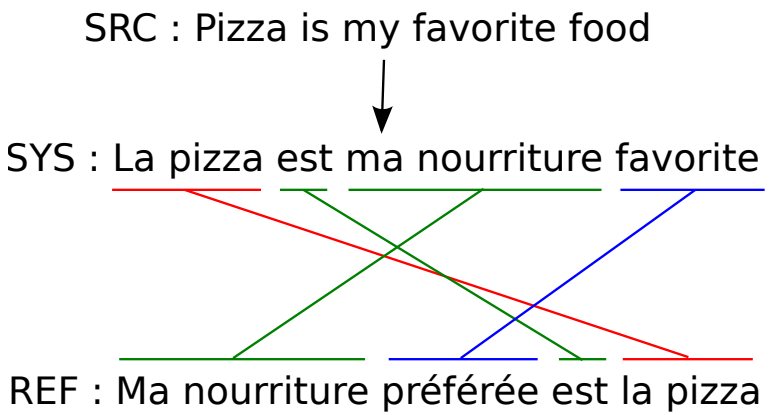
$$Rouge = \frac{\text{mots en commun}}{\text{mots de la référence}}$$

  - n résumés de référence, 1 résumé produit par le système
  - Rouge-2 : bigrammes de mots (séquences de 2 mots)
  - Rouge-SU4 : bigrammes avec un trou < 4 entre les 2 mots



Traduction automatique

- Entrées = phrase source, sorties = phrase cible
- Manuelle :
  - Compréhension
  - Fluidité
- Automatique :
  - Translation Error Rate (TER) = Nombre de transformations nécessaires



  - BLEU (Bilingual Evaluation Understudy)

$$Bleu = \frac{\text{mots en commun}}{\text{mots du système}}$$

Evaluer des systèmes interactifs

- Dialogue personne-machine
  - assistants personnels
  - serveurs téléphoniques
  - *chatbots*
- Problèmes
  - qu'est-ce qu'une *référence* humaine pour le dialogue ?
  - pas possible d'utiliser la même *référence* pour comparer des systèmes
    - ★ évaluation dynamique : chaque utilisateur aura une expérience différente
- Solutions
  - dialogue avec but (ex : réservation)
    - ★ taux de complétion de la tâche
    - ★ satisfaction utilisateur
  - conversations (chatbots)
    - ★ taux d'*engagement* des utilisateurs (temps passé)
    - ★ métriques *corrélées* : comparaison avec réponses *humaines*

Dans l'industrie

Retour sur investissement (ROI, Return on investment)

$$ROI = \frac{\text{gains} - \text{coûts}}{\text{coûts}}$$

- Ne peut être optimisé directement
- Seulement une estimation

Exemple : routage d'email automatique

- Avant : 1 personne lit et transfère les emails
- Après : les emails sont transférés automatiquement avec 10% d'erreurs
- Gains : 1 salaire de moins, plus rapide
- Coûts : prix du système + temps perdu à cause de erreurs

Exemple : détection du spam

- Avant : 90% du trafic est du spam
- Après : 10% du trafic est du spam mais 10% de bons emails étiquetés comme spam
- Gain : moins de trafic, usagés satisfaits
- Coûts : prix du système + emails perdus

Significativité

Quelle est la probabilité que Sys1 > Sys2 ?

	x	y	z	w	moy.	éq.-type
Sys1	80	72	69	45	66.5	15.1
Sys2	78	76	65	49	67.0	13.3

- Méthode du mélange stratifié (si m1 > m2) :

```
nrand = 10000
nsup = 0
sys1 = [x1, y1, z1, ...]
sys2 = [x2, y2, z2, ...]
m1 = measure(sys1); m2 = measure(sys2)
repeat nrand times:
    randomly_swap(sys1, sys2)
    n1 = measure(sys1); n2 = measure(sys2)
    if n1 - n2 >= m1 - m2:
        nsup ++
return (nsup + 1) / (nrand + 1)
```

Corrélation

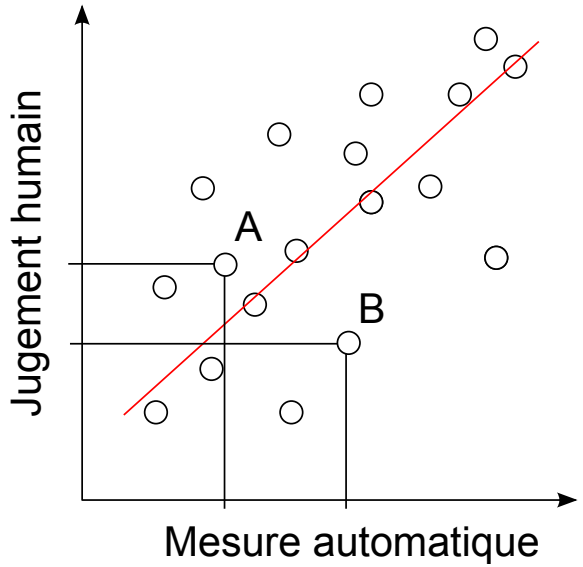
- Les annotateurs prennent-t-ils les mêmes décisions ?
- Qualité d'une mesure automatique ?

Corrélation classique :

$$corr(X,Y) = \frac{moy\left[(X - moy(X)) \times (Y - moy(Y))\right]}{eqtype(X) \times eqtype(Y)}$$

Corrélation des rangs (méthode de Kendall) :

$$\tau = \frac{\text{nb d'accord} \ - \ \text{nb pas d'accord}}{\frac{n(n-1)}{2}}$$



Bilan sur l'évaluation

- Métriques d'évaluation
  - de très nombreuses métriques
  - dépendantes de la tâche visée
  - basées sur la notion de *référence* humaine
    - ⋆ avec tous les problèmes que cela pose ! !
- Mais où trouver ces *références humaines* ?
  - **LE grand problème des méthodes d'apprentissage supervisé ! !**

Corpus + Annotation

- Corpus
  - le WEB et les documents électronique sont une source (presque) infini de corpus de texte
    - ⋆ est-ce que le WEB est réellement un *corpus* ?
      `https://www.aclweb.org/anthology/J03-3001.pdf`
    - ⋆ problème des langues sous-représentées
    - ⋆ problème des langues sans forme écrite fixe
- Annotation
  - faites par des linguistes experts
    - ⋆ cas *idéal* mais rare (car cher ! !)
  - faites par un grand nombre de gens peu qualifiés
    - ⋆ *crowd sourcing* comme *Amazon Mechanical Turk*, ou *serious games*
  - **annotations gratuites** : mythe ou réalité ?

Corpus + Annotation

- Annotations *gratuites* $\rightarrow$ faites par des utilisateurs lors de leur pratique habituelle d'un outil
  - paradigme du *click* pour la recommandation de pages internet
    - ⋆ chaque fois que vous cliquez sur un lien proposé après une requête, vous rajoutez une *annotation gratuite*
  - paradigme applicable à des tâches de TAL ? $\rightarrow$ oui, avec quelques réserves ! !
  - exemples :
    - ⋆ correction des SMS après dictée vocale
    - ⋆ commentaires avec notes sur des sites d'avis d'utilisateurs
    - ⋆ enquêtes de satisfaction après utilisation d'un service

Et maintenant, l'évaluation à l'heure du *Deep Learning* ?

- Encore plus présente ! !
  - compétition intense ! !
  - nécessité de plus en plus de corpus d'apprentissage
- Règne des évaluations *benchmarks*
  - *diktat* des fonctions objectives utilisées pour entraîner les réseaux de neurones
  - exemple : *compréhension automatique de texte*

Exemple d'applications : compréhension écrite de textes

- Questions de compréhension sur un texte

`M. Wildon a laissé plus clairement entendre que si l'Allemagne exécutait sa menace contre le commerce neutre, l'Amérique lui déclarerait la guerre et il a demandé aux neutres de se joindre à lui dans son action.`

- Questions
  - *Qui a demandé aux neutre de se joindre à lui dans son action ?*
    - ⋆ `M. Wildon`
  - *À qui est-ce que l'Amérique a déclaré la guerre ?*
    - ⋆ `l'Allemagne`
  - *Qu'est-ce que l'Amérique a déclaré à l'Allemagne ?*
    - ⋆ `la guerre`

Vastes corpus d'apprentissage pour la tâche - ex : SQuAD

Predictions by nlnet (single model) (Microsoft Research Asia)

Article EM: 72.9 F1: 76.4

Amazon_rainforest

The Stanford Question Answering Dataset

The Amazon rainforest (Portuguese: Floresta Amazônica or Amazônia; Spanish: Selva Amazónica, Amazonía or usually Amazonia; French: Forêt amazonienne; Dutch: Amazoneregenwoud), also known in English as Amazonia or the Amazon Jungle, is a moist broadleaf forest that covers most of the Amazon basin of South America. This basin encompasses 7,000,000 square kilometres (2,700,000 sq mi), of which 5,500,000 square kilometres (2,100,000 sq mi) are covered by the rainforest. This region includes territory belonging to nine nations. The majority of the forest is contained within Brazil, with 60% of the rainforest, followed by Peru with 13%, Colombia with 10%, and with minor amounts in Venezuela, Ecuador, Bolivia, Guyana, Suriname and French Guiana. States or departments in four nations contain "Amazonas" in their names. The Amazon represents over half of the planet's remaining rainforests, and comprises the largest and most biodiverse tract of tropical rainforest in the world, with an estimated 390 billion individual trees divided into 16,000 species.

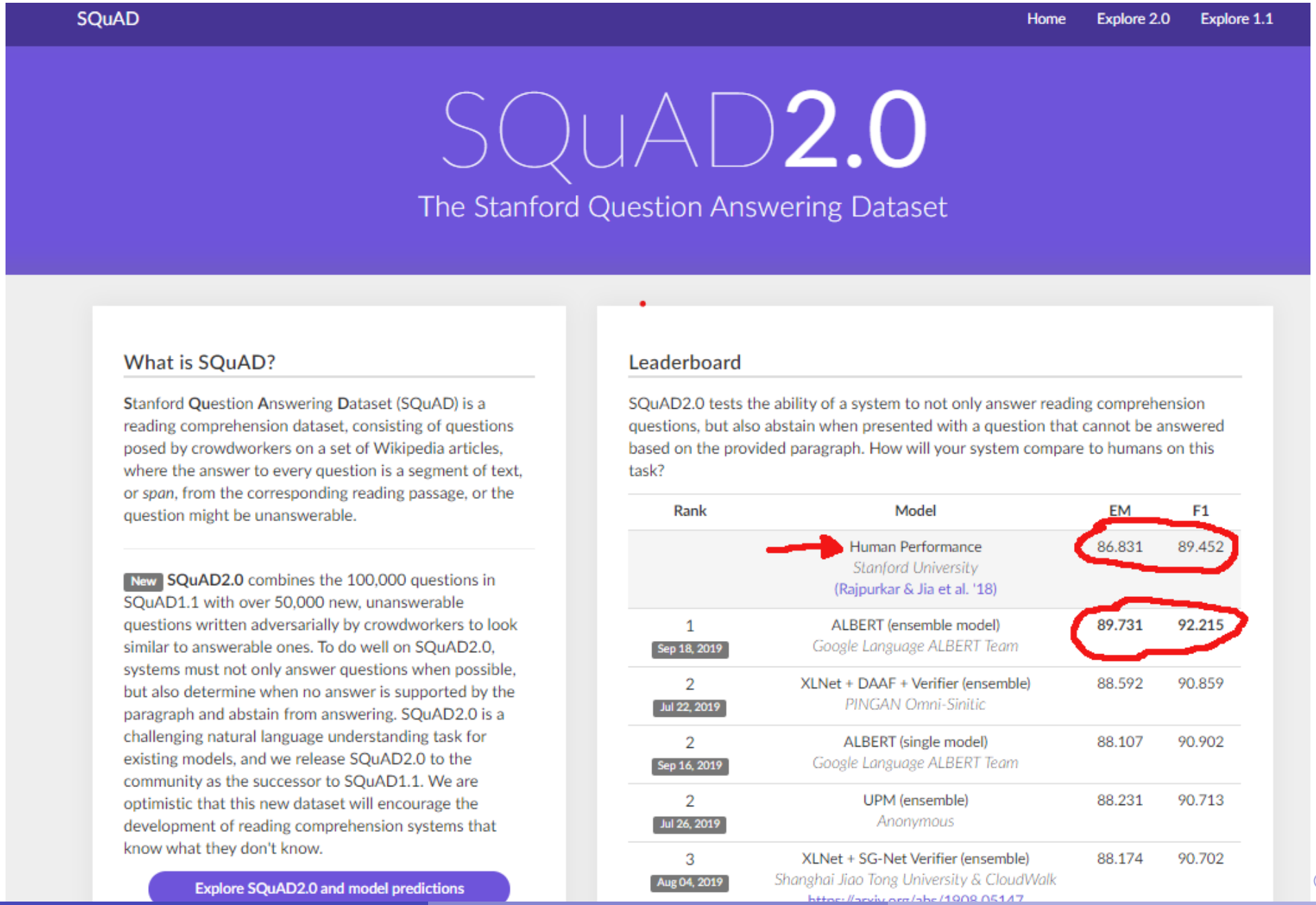
Which name is also used to describe the Amazon rainforest in English?
Ground Truth Answers: also known in English as Amazonia or the Amazon Jungle, | Amazonia or the Amazon Jungle | Amazonia
Prediction: Amazonia

How many square kilometers of rainforest is covered in the basin?
Ground Truth Answers: 5,500,000 square kilometres (2,100,000 sq mi) are covered by the rainforest. | 5,500,000 | 5,500,000
Prediction: 5,500,000

How many nations control this region in total?
Ground Truth Answers: This region includes territory belonging to nine nations. | nine | nine
Prediction: nine

How many nations contain "Amazonas" in their names?
Ground Truth Answers: States or departments in four nations contain "Amazonas" in their names. | four | four
Prediction: four

Compétition féroce et résultats brutaux



SQuAD

Home Explore 2.0 Explore 1.1

SQuAD2.0

The Stanford Question Answering Dataset

What is SQuAD?

Stanford **Q**uestion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

New **SQuAD2.0** combines the 100,000 questions in SQuAD1.1 with over 50,000 new, unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering. SQuAD2.0 is a challenging natural language understanding task for existing models, and we release SQuAD2.0 to the community as the successor to SQuAD1.1. We are optimistic that this new dataset will encourage the development of reading comprehension systems that know what they don't know.

Explore SQuAD2.0 and model predictions

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 18, 2019	ALBERT (ensemble model) *Google Language ALBERT Team*	**89.731**	**92.215**
2 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
2 Sep 16, 2019	ALBERT (single model) *Google Language ALBERT Team*	88.107	90.902
2 Jul 26, 2019	UPM (ensemble) *Anonymous*	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk*	88.174	90.702

Bilan

- Les évaluations de type *benchmark* ne rendent pas toujours compte des performances *réelles* et de l'utilité des systèmes développés
- En particulier, elles n'évaluent pas (ou mal) des capacités indispensables pour simuler un comportement humain :
  - l'introspection
  - le sens commun
  - l'adaptation
  - la remise en cause

Introspection

Comment un système peut-il percevoir qu'il est en train de faire une erreur ?

Exemple : cas de la traduction parole/parole

- **locuteur 1**
  - énoncé : *Bonjour, je voudrais savoir à qui appartiennent ces entrepôts ?*
  - transcription automatique : *Bonjour je voudrais savoir à qui appartiennent ces entre peau ?*
  - traduction automatique : *Hello I would like to know to who belongs between skin ?*
  - synthèse de signal
- **locuteur 2**
  - énoncé : *What ? ? which skin ? ? what do you mean ? ?*
  - transcription automatique : *what which skin that do you mean*
  - traduction automatique : *quelle apparence tu veux dire*
  - synthèse de signal
- **locuteur 1**
  - énoncé : *pardon ? ? non, ces entrepôts*
  - transcription automatique : *pardon non c'est en troupeau*
  - traduction automatique : *sorry no it's a flock*
  - synthèse de signal
- **locuteur 2**
  - ...

Exemple : cas de la traduction parole/parole

- **speaker 1**
  - ref. transcription : *Bonjour, je voudrais savoir à qui appartiennent ces entrepôts ?*
  - ASR transcription : *Bonjour je voudrais savoir à qui appartiennent ces entre peau ?*
- **system BOLT**
  - speech synthesis : *Excusez-moi, je ne connais pas ce mot, à qui appartiennent quoi ?*
- **speaker 1**
  - ref. transcription : *ces hangars de l'autre côté de la rue*
  - ASR transcription : *ces hangars de l'autre côté de la rue*
- **system BOLT**
  - ref. transcription : *OK, je traduis*
  - ASR transcription clarified : *Bonjour je voudrais savoir à qui appartiennent ces hangars de l'autre côté de la rue ?*
  - automatic translation : *Hello I would like to know who owns these warehouse on the other side of the street ?*
  - speech synthesis
- **speaker 2**
  - ...

Sens commun

Comment un système peut-il intégrer des connaissances générales sur le monde ?

Sens commun

- `La statue ne rentre pas dans la boîte car elle est trop grande`
  - Qui est trop grand ?
- `Jean lâche le ballon`
  - Où est le ballon ?

Adaptation

Comment un système peut-il adapter ses modèles pour pouvoir traiter une nouvelle application ?

Remise en cause

Comment un système peut-il apprendre de ses erreurs ?