Mots de passe : le choix humain plus sécurisé que la génération aléatoire

Nicolas K. Blanchard [1], Clément Malaingre [2], Ted Selker[3]

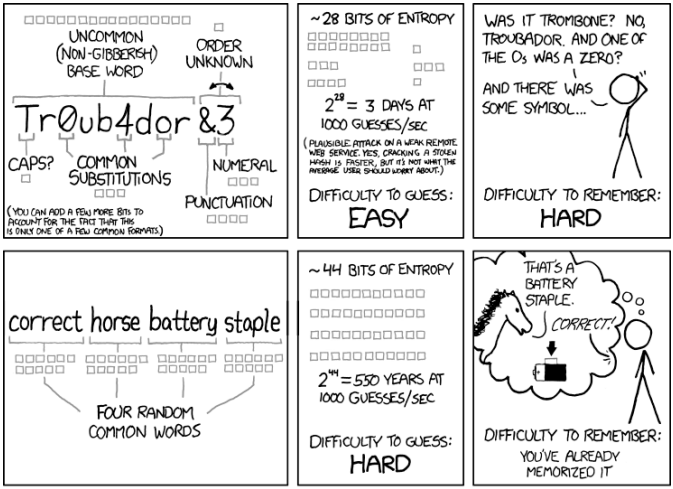
[1]IRIF, Université Paris Diderot

[2]Teads France

[3]University of California, Berkeley

20èmes Rencontres Francophones pour les Aspects Algorithmiques des Télécommunications 31 mai 2018

Phrases de passe versus mots de passe (xkcd)



Phrases de passe versus mots de passe

Avantages des phrases de passe grâce à l'utilisation du langage humain :

- Facilité de mémorisation supérieure
- Plus grande entropie (à mémorisation équivalente)
- Utiles comme primitive pour certains algorithmes de sécurité

Méthodes actuelles de création

Première possibilité : laisser les personnes créer leurs propres phrases.

Méthodes actuelles de création

Première possibilité : laisser les personnes créer leurs propres phrases.

Problèmes :

- Utilisation de phrases de la littérature (chansons/poèmes)
- Utilisation de phrases connues (jusqu'à 2.55% d'utilisateurs ayant la même phrase dans une expérience)
- Phrases de basse entropie avec des mots courants

Méthodes actuelles de création

Première possibilité : laisser les personnes créer leurs propres phrases.

Problèmes :

- Utilisation de phrases de la littérature (chansons/poèmes)
- Utilisation de phrases connues (jusqu'à 2.55% d'utilisateurs ayant la même phrase dans une expérience)
- Phrases de basse entropie avec des mots courants

Deuxième possibilité : génération aléatoire.

Méthodes actuelles de création

Première possibilité : laisser les personnes créer leurs propres phrases.

Problèmes :

- Utilisation de phrases de la littérature (chansons/poèmes)
- Utilisation de phrases connues (jusqu'à 2.55% d'utilisateurs ayant la même phrase dans une expérience)
- Phrases de basse entropie avec des mots courants

Deuxième possibilité : génération aléatoire.

Limites :

- Petit dictionnaire si on veut que la personne connaisse les mots
- Mémorisation plus difficile

Que se passe-t-il en hybridant les deux méthodes ?

Expérience utilisateur

Expérience simple : on montre une liste de 20 ou 100 mots à des utilisateurs, qui en choisissent 6.

Questions :

- Quels sont les facteurs qui influencent le choix des mots ?

Expérience utilisateur

Expérience simple : on montre une liste de 20 ou 100 mots à des utilisateurs, qui en choisissent 6.

Questions :

- Quels sont les facteurs qui influencent le choix des mots ?
- Comment cela affecte-t-il l'entropie ?

Expérience utilisateur

Expérience simple : on montre une liste de 20 ou 100 mots à des utilisateurs, qui en choisissent 6.

Questions :

- Quels sont les facteurs qui influencent le choix des mots ?
- Comment cela affecte-t-il l'entropie ?
- Quelles sont les erreurs les plus fréquentes ?

Expérience utilisateur

Expérience simple : on montre une liste de 20 ou 100 mots à des utilisateurs, qui en choisissent 6.

Questions :

- Quels sont les facteurs qui influencent le choix des mots ?
- Comment cela affecte-t-il l'entropie ?
- Quelles sont les erreurs les plus fréquentes ?
- Gagne-t-on en mémorisation ?

Protocole

Protocole simple :

- Montrer une liste de 20 ou 100 mots pris uniformément au hasard à chaque utilisatrice
- Demander de choisir et d'écrire une séquence de 6 mots (ou imposer la séquence de 6 mots au groupe contrôle)
- Demander aux utilisatrices de mémoriser la séquence et donner un exercice pour aider
- Distraire en montrant une deuxième liste et en demandant de "deviner" le choix d'une autre participante
- Demander de réécrire la séquence initiale

Interface

homogenization	parabolic	hydride	refits	piezometer
passe	pralines	radicalised	sanctuaries	ejecting
erotically	wickets	sperm	almandine	devourer
cenotes	pointedness	noninfectious	enhances	tenterhooks
turned	microtonal	chimaera	underwrite	upturns
colorations	hayrides	symbolical	relinquished	above
scant	invulnerable	reservations	sophistry	paramyxovirus
camphor	incalculable	novena	biomaterials	turn
samaritans	supercontinent	touchy	divvied	speeds
freewheel	translocates	bioinformatics	ants	attractiveness
relocation	antioxidants	spears	respected	vernaculars
fuhrer	moribund	incapacitating	apolipoproteins	kalis
myocarditis	resignedly	redesigns	physiology	pinewood
sulky	silky	retrogressive	backword	rhapsody
talpa	memorialize	hazard	keynoter	masons
disown	fermion	endowment	semifinalist	cards
subsumption	serendipitous	molla	housemaids	coach
potter	quandary	mod	kores	downlight
treehouse	off	mib	bayle	desexed
chinese	planetesimal	chapbook	kale	pyrophosphate

Submit

Démographie

Statistiques démographiques :

- 99 participants en deux groupes (plus 26 de groupe contrôle)
- Entre 16 et 69 ans, moyenne de 31 ans.
- Majorité anglophone, un quart francophone, un sixième hébréophone

Modèles de choix

Uniforme : chaque mot avec probabilité égale

Modèles de choix

Uniforme : chaque mot avec probabilité égale

Plus Fréquents : uniquement les six mots les plus fréquents de la liste

Uniforme : chaque mot avec probabilité égale

Plus Fréquents : uniquement les six mots les plus fréquents de la liste

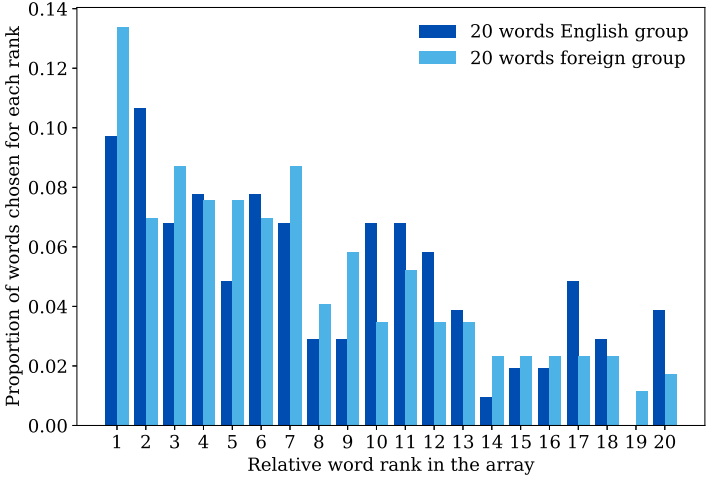
Corpus : chaque mot en probabilité proportionnelle à son biais naturel dans un corpus littéraire (renormalisé). Mot de rang r_k pris dans une liste de n mots avec probabilité :

$$\frac{\frac{1}{r_k}}{\sum_{i=1}^{n} \frac{1}{r_i}}$$

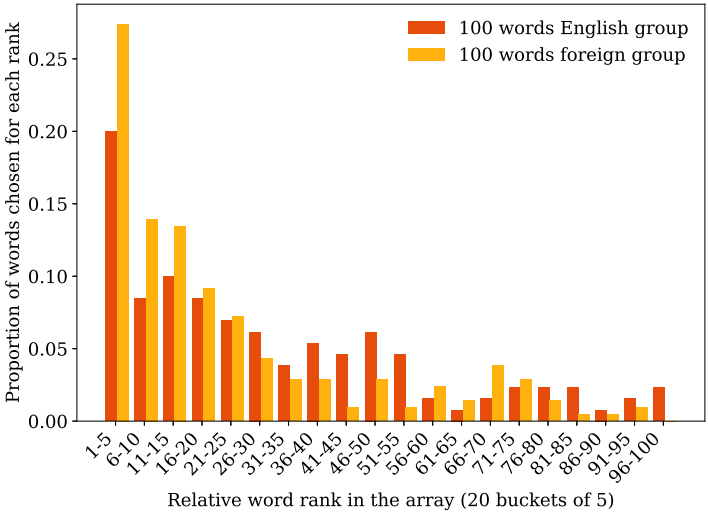
Comparaisons des erreurs

Essai	Correct	Manquant	Faux	Typo	Variante	Ordre
1:20	18/47	26	5	6	8	6
1:100	26/51	16	4	10	5	3
Contrôle	6/26	31	12	11	11	10
2:20	14/29	0	3	1	2	8
2:100	15/26	1	4	4	2	3

Biais positionnel



Biais positionnel



Biais sémantique

72.6	53.5	47.8	47.8	40.1	137
47.8	45.9	24.8	32.5	22.9	91
32.5	24.8	19.1	11.5	11.5	52
13.4	17.2	13.4	7.6	13.4	34
87	74	55	52	46	314

Biais sémantique

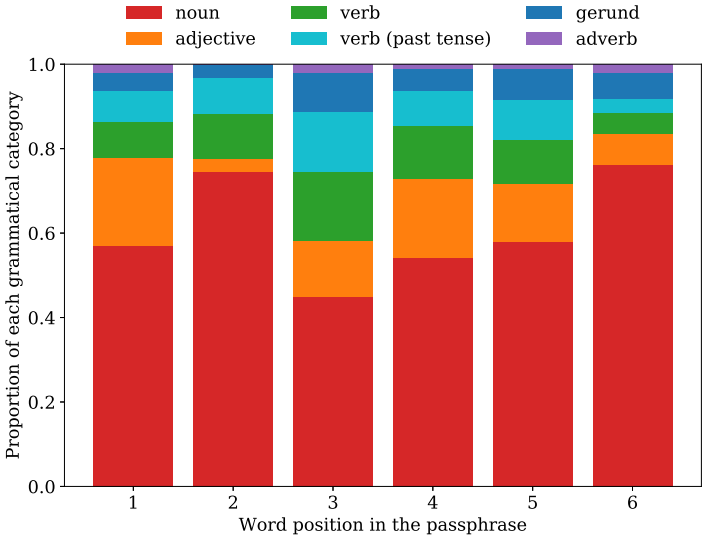
32.6	17.9	29.3	31.0	19.6	80
11.4	13.0	4.9	9.8	16.3	34
6.5	17.9	14.7	17.9	14.7	44
4.9	8.2	9.8	11.4	13.0	29
6.5	11.4	8.2	9.8	6.5	26
6.5	6.5	1.6	9.8	8.2	20
6.5	8.2	1.6	6.5	1.6	15
6.5	0.0	6.5	11.4	1.6	16
3.3	0.0	3.3	4.9	8.2	12
4.9	3.3	9.8	1.6	6.5	16
8.2	1.6	1.6	6.5	1.6	12
0.0	3.3	4.9	1.6	4.9	9
1.6	6.5	0.0	1.6	1.6	7
3.3	1.6	3.3	6.5	3.3	11
6.5	1.6	4.9	3.3	0.0	10
1.6	0.0	1.6	1.6	4.9	6
1.6	3.3	1.6	0.0	3.3	6
1.6	1.6	1.6	0.0	1.6	4
1.6	1.6	1.6	3.3	0.0	5
1.6	1.6	1.6	1.6	3.3	6
72	67	69	86	74	368

Biais syntaxique

Effets syntaxiques :

- Phrases ayant un sens modérément fréquentes ($< 50\%$)
- 65 structures différentes pour 99 phrases
- Une seule structure présente plus de trois fois : 6 noms d'affilée

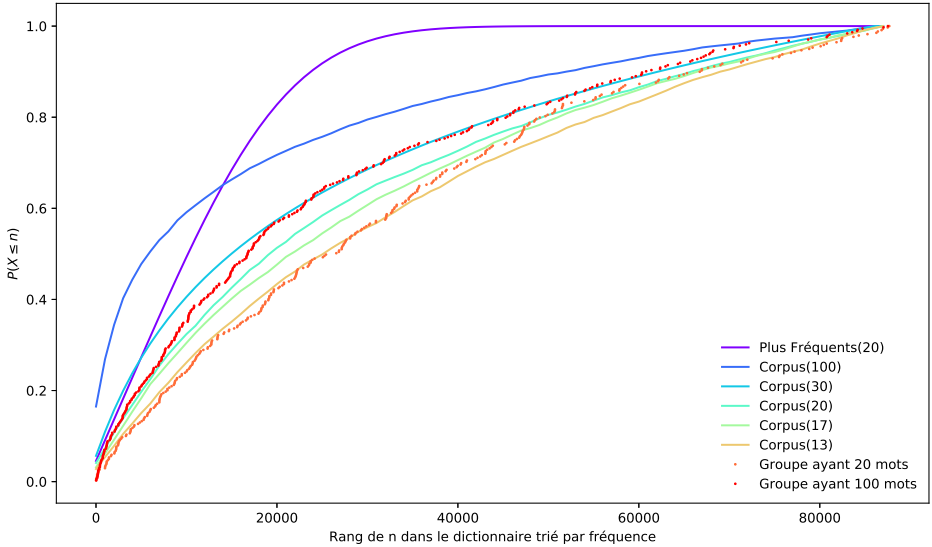
Biais syntaxique



Comparaison d'entropies

Stratégie	Entropie (bits)
Uniforme(dictionnaire de 87691 mots)	16.42
Corpus(13)	16.25
Corpus(17)	16.15
Corpus(20)	16.10
Corpus(30)	15.91
Corpus(100)	15.32
Plus Fréquents(20)	12.55
Uniforme(5000)	12.29
Plus Fréquents(100)	10.69
Corpus(langue anglaise : 276663 mots)	8.94
Corpus(dictionnaire de 87691 mots)	8.20

Courbes d'entropies



Conclusion

Conclusion

Avantages de la méthode avec 100 mots :

- Sécurisée : 97% de l'entropie maximale, soit 30% de plus qu'avec un dictionnaire limité
- Mémorable : division par quatre du taux d'erreur du groupe contrôle
- Transférable : outil peut s'exécuter dans le navigateur du client, pesant moins de 1Mo

Conclusion

Avantages de la méthode avec 100 mots :

- Sécurisée : 97% de l'entropie maximale, soit 30% de plus qu'avec un dictionnaire limité
- Mémorable : division par quatre du taux d'erreur du groupe contrôle
- Transférable : outil peut s'exécuter dans le navigateur du client, pesant moins de 1Mo

Questions :

- Quel est le nombre optimal de mots à montrer ?
- Peut-on rendre cela plus uniforme en utilisant un nuage de tag ou des méthodes similaires ?
- Cela vaut-il le coup de prendre des dictionnaires encore plus complets ?
- Comment créer un modèle complet du choix humain intégrant les différentes variables ?

Merci pour votre attention