L'évaluation des systèmes de RI

L. Maisonnasse

Pourquoi ?

- Beaucoup de variations dans les méthodes de RI
  - Modèles
    - Vectoriel, probabiliste, modèle de langue
    - Pondérations des termes (TF, TF-IDF,...)
  - Algorithmes
    - Correspondance (produit cartésien, cosinus, ...)
    - Retour de pertinence, expansion de requêtes
  - Systèmes
    - Sélection des termes (suppression des mots-outils, racinisation...)
- Comment trouver le meilleur système, les meilleurs composants ?
  - Comparer les SRI de manière théorique (via leur modèle)
    - Problème non-résolu, pas de modèle générique
  - Tester les résultats d'un système
    - Evaluations de type « boîte noire »

Que tester ?

- Un système de recherche d'information doit satisfaire un besoin d'information d'un utilisateur

pertinence utilisateur ≠ pertinence système

- Pertinence utilisateur : satisfaction de l'utilisateur
- Pertinence système : estimation du système

But de l'évaluation : comparer les deux

3

Difficultés de l'évaluation

- Les performances dépendent des objets retrouvés
  - Quelle partie doit-on retrouver ?
- La pertinence n'est pas binaire mais plutôt continue.
  - Même binaire il est parfois difficile de juger un objet
- Pour un humain la pertinence est :
  - Subjective : dépend de l'utilisateur
  - Situationnelle : dépend de la situation
  - Cognitive : dépend de la perception et du comportement
  - Dynamique : évolue avec le temps

4

Standardiser le processus : corpus d'évaluation

- Définir une collection de documents
  - Réaliste
  - Couverture de problèmes
- Définir un ensemble de requêtes sur ce corpus
  - Réaliste
  - Evaluable
- Déterminer exhaustivement les documents pertinents pour chaque requête
  - Humainement
  - Pertinence binaire

➔ Effort de création très grand

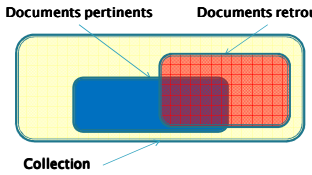
5

Les mesures

Comparer la pertinence système et la pertinence utilisateur pour une requête

Rappel et précision

- Pour une requête

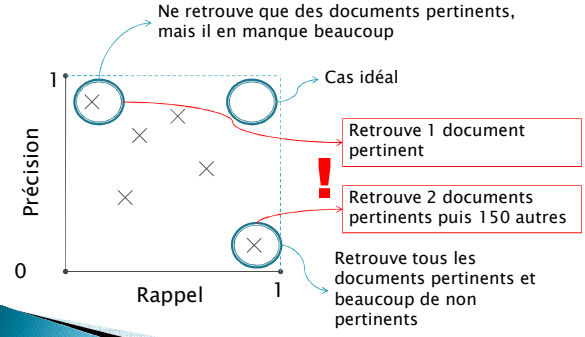


Pertinent \ Retrouvé	oui	non
non	Retrouvé et non pertinent	Non retrouvé et non pertinent
oui	Retrouvé et pertinent	Non retrouvé et pertinent

$$rappel = \frac{Nombre\ de\ documents\ pertinents\ retrouvés}{Total\ de\ documents\ pertinents}$$

$$précision = \frac{Nombre\ de\ documents\ pertinents\ retrouvés}{Total\ de\ documents\ retrouvés}$$

Comparer le rappel et la précision



8

Calculer des points de rappel précision

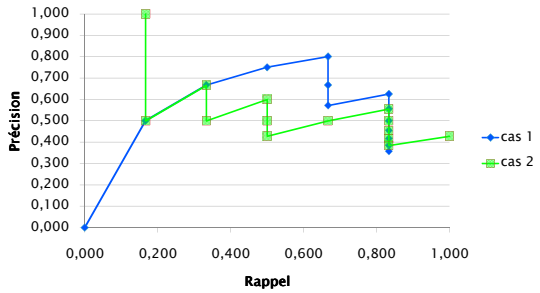
- Une requête donne une liste ordonnée de documents
  - Plusieurs calculs de rappel et précision possibles
    - A chaque sélection des N premiers documents
  - Comparer la liste des documents
- Méthode
  - Marquer les documents pertinents de la liste
  - Calculer le rappel et la précision pour chaque position dans la liste
  - Positionner les points sur une courbe rappel/précision

9

Exemple avec deux variations

	Cas 1					Cas 2				
n	doc	pert	#pert	rappel	précision	doc	pert	#pert	rappel	précision
1	588		0	0,000	0,000	588	x	1	0,167	1,000
2	589	x	1	0,167	0,500	576		1	0,167	0,500
3	576	x	2	0,333	0,667	589	x	2	0,333	0,667
4	590	x	3	0,500	0,750	342		2	0,333	0,500
5	986	x	4	0,667	0,800	590	x	3	0,500	0,600
6	592		4	0,667	0,667	717		3	0,500	0,500
7	984		4	0,667	0,571	984		3	0,500	0,429
8	988	x	5	0,833	0,625	772	x	4	0,667	0,500
9	578		5	0,833	0,556	321	x	5	0,833	0,556
10	985		5	0,833	0,500	498		5	0,833	0,500
11	103		5	0,833	0,455	113		5	0,833	0,455
12	591		5	0,833	0,417	628		5	0,833	0,417
13	772		5	0,833	0,385	772		5	0,833	0,385
14	990		5	0,83333	0,35714	592	x	6	1	0,42857

Sous forme graphique

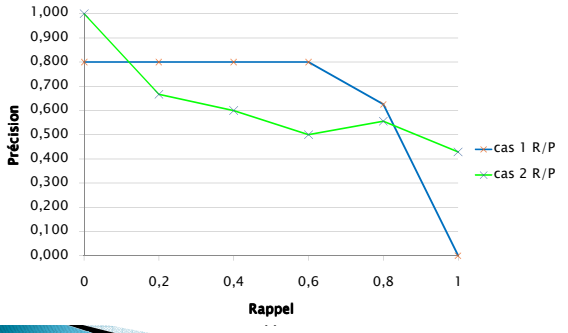


Peu lisible : il faut interpoler

- Déterminer la précision à des points standards de rappel
  - 6 points de rappel
    - {0.0, 0.2, 0.4, 0.6, 0.8, 1.0}
  - 11 points de rappel
    - {0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}
- Interpoler :
  - La précision à un point de rappel est la précision maximum après ce point

$$P(r_i) = \max_{i<j} P(r_j)$$

Courbe de rappel/précision avec 6 points



13

Les autres mesures

- Rappel et précision
  - Plusieurs mesures possibles
  - Difficile de comparer
    - Les courbes se croisent
    - Les courbes sont proches
  - Signification trop générale
- Solution
  - Synthétiser
    - Une mesure unique
    - Sur plusieurs requêtes : faire la moyenne
  - Avec du sens
    - Mesure orientée

Précision moyenne

- Précision moyenne (Par requête)
  - Moyenne des valeurs de précision pour chaque point de rappel (6/10/11)
    - Ex1: (0,8+0,8+0,8+0,8+0,625+0)/6 = 0,638
    - Ex2: (1+0,667+0,6+0,5+0,556+0,429)/6 = 0.625
  - Interpoler ou non …
- Précision moyenne (Pour le système)
  - Moyenne des précisions moyennes de chaque requête
- Vision globale des performances

Précision à N

- Précision pour les N premiers documents retrouvés
- Orientée précision si N faible
- Exemple
  - Précision à 3 docs
    - 0,667
  - Précision à 5 docs
    - 0.8

n	doc	pert	#pert	rappel	précision
1	588		0	0,000	0,000
2	589	x	1	0,167	0,500
3	576	x	2	0,333	0,667
4	590	x	3	0,500	0,750
5	986	x	4	0,667	0,800
6	592		4	0,667	0,667
7	984		4	0,667	0,571
8	988	x	5	0,833	0,625
9	578		5	0,833	0,556

16

F-Score

- Une mesure qui effectue une balance entre le rappel et la précision
- Moyenne harmonique F-Score
  - Les deux doivent être élevés

$$F_{score} = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

- Moyenne pondérée F-Score

$$F_{\beta} = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R}+\frac{1}{P}}$$

  - β=1 ➔ F_{score}
  - B>1 ➔ Favorise le rappel
  - B<1 ➔ Favorise la précision
- Problème : Quel rappel choisir ?

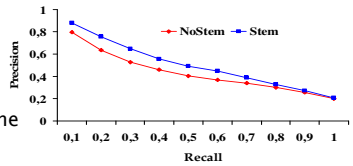
17

Sur un ensemble de requêtes

- Pourquoi la moyenne ?
  - Utilisation de la GMAP
    - Favorise les systèmes robustes
- La différence est-elle suffisante ?
  - Différences communément admises
    - Différence de 5% sur la MAP
  - Comparer les résultats sur une liste de requêtes
    - La différence entre deux systèmes a-t-elle un sens ?
    - Peu de requêtes différentes
  - Test de significativité
    - T-test
    - Sign test
    - Wilcoxon test

Comparer deux systèmes

- Les mêmes données
- La courbe rappel précision
  - (en haut à droite)



- La précision moyenne
  - Performance globale
- Mesure spécifique
  - Premiers documents retrouvés Précision à 5 documents
- Calcul de la significativité

Limites des mesures

- Problèmes de la précision et du rappel
  - Le nombre de documents non pertinents dans la collection
  - Le rappel est indéfini quand aucun document pertinent n'est retrouvé
  - La précision est indéfinie quand aucun document n'est retrouvé.
- Problèmes de la pertinence
  - Diversité
  - Nouveauté
- Autres facteurs
  - L'effort de l'utilisateur
  - Temps de réponse
  - Interface

  ➔ Temps de recherche

Le protocole expérimental

Collections d'évaluation

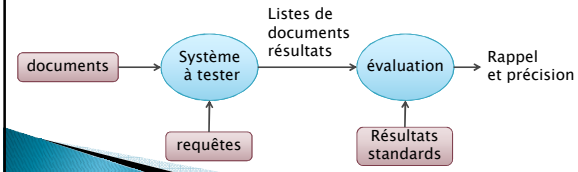
Les objectifs

- Donner des bases communes pour comparer différents systèmes de RI
  - Mêmes documents, requêtes, mesures.
- Proposer de grandes collections
- Proposer des méthodes et des mesures spécifiques

22

Collection d'évaluation

- Une collection d'évaluation
  - Un ensemble de documents (réutilisable)
  - Un ensemble de requêtes, de besoin d'information
  - Une liste des documents pertinents pour chaque requête
  - Des mesures appropriées
- Processus



23

Les difficultés

- Les performances ne sont valides que pour une collection d'évaluation
  - Manque de généralité
- La construction est difficile et couteuse
  - Recherche d'un corpus réaliste, choix des documents
    - Représentativité par rapport à la tâche
    - Diversité des sujets, du vocabulaire
    - Texte intégral vs. résumé
  - Annotation de documents pertinents
- Adapter les mesures au type de tâche
  - La pertinence n'est pas toujours simple
- Définir la quantité suffisante
  - Quels et combien de besoins d'information ?
    - Le nombre doit être supérieur à 25

24

Les premières collections

- Collections SMART (ftp://ftp.cs.cornell.edu/pub/smart)
  - Différents domaines
    - cisi, domaine des sciences de l'information.
    - med, domaine médical.
    - time, dépêches du times.

collection	taille	nombre de documents	nombre de requêtes	contenu des documents			
adi	36 K	82	35	titre	auteurs	résumé	
cacm	2.1M	3204	64	titre	auteurs	citations	
cisi	2.3M	1460	112	titre	auteurs	résumé	citations
cran	1.6M	1400	225	titre	auteurs	résumé	
med	1.0M	1033	30	résumé			
npl	3.1M	11429	93	titre	long		
time	1.5M	425	83	dépêche			

- Utiliser par différents chercheurs pour tester leurs méthodes
  - Trop petites

25

L'évaluation TREC

- TREC: Text REtrieval Conference (http://trec.nist.gov/)
  - Organiser par le programme TIPSTER et supporté par la DARPA (Defense Advanced Research Projects Agency).
- Conférence annuelle d'évaluation depuis 1992
  - NIST (National Institute of Standards and Technology) et DARPA.
- Favoriser les participations académique et industrielle
- Développer des nouvelles techniques d'évaluation
  - Notamment pour de nouvelles applications
- Varier les collections (pistes) pour évaluer différents aspects de la RI
  - Recherche, routage/filtrage, collection non anglaise, collection web, question réponse.
- Travailler sur de grandes collections
  - Recherche : 500 000 documents
  - Web : 1To de données

26

La méthode TREC

- Chaque année des documents sont fournis avec des besoins à résoudre
  - Fournit également du matériel d'apprentissage
- Les participants soumettent les listes de documents retrouvées pour chaque requête
  - Définit les documents pertinents
  - Calcule les différentes mesures
- Les participants présentent leurs résultats lors de la conférence
  - Décrit des méthodes, des mesures

=> Déterminer les meilleures approches

Le problème majeur : trop de documents à évaluer

- Exemple:
  - Déterminer tous les documents pertinents sur une collection de 500 000 documents pour une requête
  - Un évaluateur très rapide prend 10s pour évaluer un document
  - Temps d'évaluation de tous les documents 10s x 500 000 = 5 000 000s ~ 58 Jours pleins
- La solution : ne pas tout évaluer
  - On crée un ensemble (Pool) des documents potentiellement pertinents
  - On utilise les 100 premiers documents de chaque participant
  - L'évaluateur ne parcourt que les documents du pool pour une requête
    - Les documents hors du pool ne sont pas examinés.
- On obtient une estimation du rappel (Zobel,1998)
- Exemple TREC-8, pour 71 participants:
  - 7 100 documents dans le pool
  - 1 736 documents uniques (Filtrage des doublons) 10s x 1 736 = 17 360 ~ 5h
  - 94 documents jugés pertinents

Bilan du modèle TREC

- Très grandes collections
- Les collections peuvent être réutilisées (outil trec_eval)
  - Rappel non complet
- Participation importante
  - TREC 1: 28 papiers 360 pages
  - TREC 4: 37 papiers 560 pages
  - TREC 7: 61 papiers 600 pages
  - TREC 8: 74 papiers
- Problème de sur-apprentissage
  - Les meilleurs résultats ne sont pas obligatoirement innovants
  - Déséquilibre des capacités …

29

Les campagnes d'évaluation et leurs pistes

- TREC
- CLEF (Cross Language Information Retrieval)
  - Créée en 2000, et destinée à l'évaluation de RI multilingue
- NTCIR
  - Depuis 1997 sur documents en langues asiatiques
- INEX
  - Lancée en 2002 pour la RI sur des documents structurés en XML.

TREC- adhoc (1/2)

- Des requêtes longues ou courtes

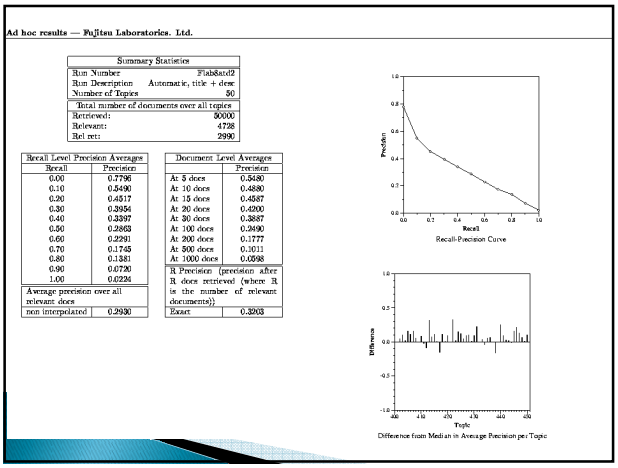
```
<top>
<num> Number: 451
<title> What is a Bengals cat?
<desc> Description:
Provide information on the Bengal cat breed.
<narr> Narrative:
Item should include any information on the Bengal cat breed,
including description, origin, characteristics, breeding program,...
</top>
```

- Collection
  - Documents bruts
    - Pas de correction du texte
  - Exemple

WSJ	Wall Street Journal articles (1986-1992)	550 Mo
AP	Associate Press Newswire (1989)	514 Mo
ZIFF	Computer Select Disks (Ziff-Davis Publishing)	493 Mo
FR	Federal Register	469 Mo
DOE	Abstracts from Department of Energy reports	190 Mo

TREC – adhoc (1/2)

- Evaluation
  - Statistiques : Nombre de besoins d'information, de documents retrouvés, de documents pertinents retrouvés
  - Rappel précision moyen
    - Moyenne sur les requêtes de 11 points de rappel
  - Précision à différents niveaux
    - 5, 10, .., 100, … 1000 documents
  - Histogramme des précisions de chaque requête
    - Différence de la précision avec les moyennes des autres participants
- Fourni par l'outil : Trec_eval



Ad hoc results — Fujitsu Laboratories. Ltd.

Summary Statistics	
Run Number	FlabSstd2
Run Description	Automatic, title + desc
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4728
Rel ret:	2990

Recall Level Precision Averages	
Recall	Precision
0.00	0.7796
0.10	0.5490
0.20	0.4517
0.30	0.3954
0.40	0.3397
0.50	0.2893
0.60	0.2291
0.70	0.1745
0.80	0.1381
0.90	0.0720
1.00	0.0226
Average precision over all relevant docs	
non interpolated	0.2930

Document Level Averages	
	Precision
At 5 docs	0.5480
At 10 docs	0.4880
At 15 docs	0.4587
At 20 docs	0.4200
At 30 docs	0.3887
At 100 docs	0.2490
At 200 docs	0.1777
At 500 docs	0.1011
At 1000 docs	0.0598
R Precision (precision after R docs retrieved (where R is the number of relevant documents))	
Exact	0.3203

TREC autres : BLOGs

- Explorer les comportements de recherche dans la blogosphère
- Collection (instantanée de la blogosphère)
  - Flux, liens permanents associés, pages d'accueil

	nombre	taille
Flux	753,681	38.6GB
Liens permanents	3,215,171	88.8GB
Pages d'acceuil	324,880	20.8GB

  - Prise en compte des opinions
    - Réordonnancer les résultats
    - Création de deux listes

Relevance	Label	# of Docs	%
Not Judged	-1	0	0%
Not Relevant	0	47491	70.5%
Relevant	1	8361	12.4%
Negative	2	3707	5.5%
Mixed	3	3664	5.4%
Positive	4	4159	6.2%
(Total)	-	67382	100%

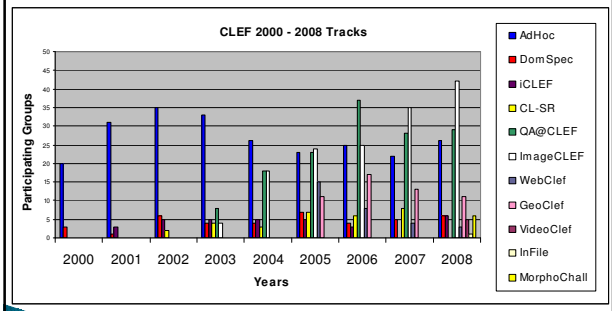
TREC autre (2/2)

- Million Query
  - Tester le rapport entre nombre de requêtes et nombre de documents à évaluer
  - Protocole
    - 10 000 requêtes à traiter, environ 2 000 évaluées
    - 25 000 000 pages web
    - Méthode de sélection des documents à évaluer
- Cross-Language Track (CLEF)
- SPAM Track
  - Tester les méthodes de filtrage
- Video Track (TRECVID depuis 2003)
  - Recherche de vidéos segmentées automatiquement, basée sur le contenu
- Beaucoup d'autres pistes :
  - Genomics / Interactive / Robust Retrieval / Relevance Feedback / Filtering / Web / Terabyte / Question Answering / HARD / Novelty /Enterprise / Legal

INEX

- Recherche d'information dans les documents structurés
  - Trouver la bonne granularité
- Corpus
  - Wikipédia (Xml) 4.6GB
- Mesures spéciales
  - Prise en compte des Doxels
    - Précision généralisée
    - Rappel généralisé

CLEF (1/2)



CLEF (2/2)

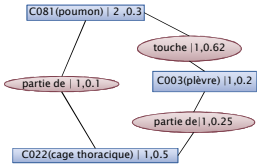
- AdHoc
  - D'une langue à une autre
  - Beaucoup de couples de langues
- QA
  - Répondre à une question dans une langue à l'aide d'information dans différentes langues
- Image
  - But
    - Retrouver une image à partir de requêtes dans différentes langues
    - Images décrites dans plusieurs langues
    - Annotation d'images / recherche
  - Deux domaines
    - Médical
    - Général (tourisme)

Quelques Expériences à CLEF images médicales

- Modèle langue à base de graphes de concepts

$$P(G_Q|M_D^g) = P(C_Q|M_D^g) \times P(R_Q|C_Q, M_D^g)$$

- Graphe de concepts
  - Détectés sur le texte
  - Utilise une base de connaissances



- Possibilité de combiner différents graphes
  - Trois méthodes de détection des graphes

Les données

- Collection CLEF médicale (2005-07)
  - Recherche d'images médicales
  - 85 requêtes

Show me x-ray images with fractures of the femur.
Zeige mir Röntgenbilder mit Brüchen des Oberschenkelknochens.
Montre-moi des fractures du fémur.

  - 50 000 documents

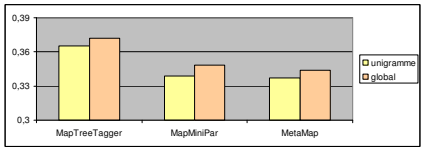
Collection	Cas	Images	Annotations	Annotations par langue
Casimage	2076	8725	2076	Français - 1899 Anglais - 177
MIR	407	1177	407	Anglais - 407
PEIR	32319	32319	32319	Anglais - 32319
PathoPic	7805	7805	15610	Allemand – 7805 Anglais - 7805
MyPACS	3577	15140	3577	Anglais - 3577
CORI	1496	1496	1496	Anglais - 1496
Total	47680	66662	55485	

Evaluation

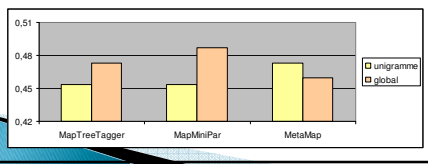
- But
  - Évaluer la recherche textuelle à base de graphes
  - Sur le texte…
- Création du graphe Ressources UMLS
  - 1 million de concepts pour 5 millions de termes
  - 54 relations sémantiques
- Mesures
  - Précision moyenne
  - Précision à 5 documents
  - Participation à CLEF 2007 et 2008

Les expériences : modèle simple

- Précision moyenne Collection CLEF médicale 2005 et 2006

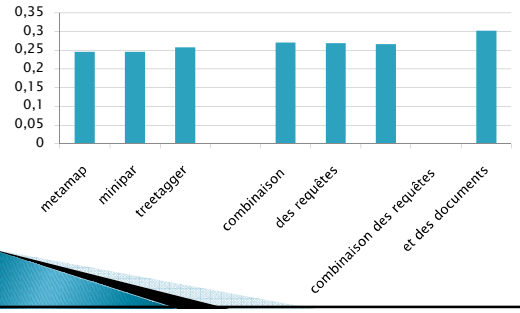


- Précision à 5 documents Collection CLEF médicale 2005 et 2006



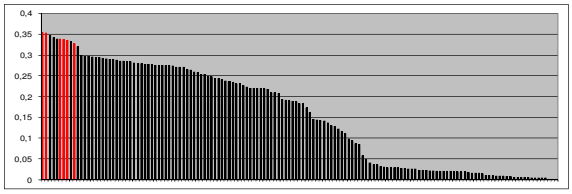
Les expériences : combinaison de modèles

- Précision moyenne Collection CLEF médicale 2005-06-07



Participation à CLEF 2007

- X participants
  - 147 runs image et/ou texte



Bilan

- La pertinence reste difficile a évaluer
  - Mais émergence de consensus
  - Évolue et s'adapte aux tâches
- Essentielle dans le domaine RI
  - Les modèles sont trop variables
  - Facilite la validation des systèmes (protocole expérimental)
- Forte volonté dans le domaine
  - Participation de plus en plus élevée
  - Nécessaire aux publications
- Risque de *formatage* des méthodes

Références

- Ellen M. Voorhees and Donna Harman, TREC Experiment and Evaluation in Information Retrieval. MIT Press, 2005.

- TREC NIST website: http://trec.nist.gov
- INEX website http://inex.is.informatik.uni-duisburg.de/2007/
- CLEF website http://clef-campaign.org/