Chapitre 3 Présentation des corpus utilisés

Ce chapitre présente les corpus auxquels nous ferons référence dans la suite de ce mémoire. Deux corpus ont été principalement étudiés : le corpus Reuters-21578 et le corpus de la tâche de filtrage de TREC-8. Le but de cette présentation est de mettre en évidence les caractéristiques de chacun de ces corpus ainsi que leurs différences. Le corpus Reuters-21578 a notamment une taille beaucoup plus réduite que le corpus de TREC-8 et présente l'avantage de regrouper des thèmes avec beaucoup de documents pertinents et d'autres pour lesquels il y en a peu.

Le corpus utilisé pour la tâche de filtrage de TREC-9 est présenté à l'annexe A car il n'est que dans le chapitre 8.

3.1 Le corpus Reuters-21578

Le corpus Reuters-21578 est un ensemble de dépêches financières émises au cours de l'année 1987 par l'agence Reuters, en langue anglaise, et disponible gratuitement sur le web[1]. Ce corpus est une mise à jour du corpus Reuters-22173 étudié notamment par [Lewis, 1992], [Wiener, 1993], [Moulinier, 1997]. Cette mise à jour, effectuée en 1996, a permis de supprimer les documents présents deux fois, de corriger des erreurs typographiques, de préciser certains formats, et de mieux définir le découpage à considérer pour l'apprentissage et le test. Du fait de ces corrections, il n'est pas possible de comparer les performances obtenues sur les différentes versions du corpus. Cependant, les caractéristiques globales du corpus sont restées identiques, et les remarques sur le comportement général des systèmes étudiés sont toujours valables.

La Figure 3.1 est un exemple de texte présent dans le corpus ; on peut noter la présence des sigles *<AXP>* et *<MER>* utilisés pour signaler des noms d'entreprise. Cet exemple montre qu'il s'agit de textes de style journalistique, rédigés.

[1] http://www.research.att.com/~lewis/reuters21578.html

Dans toute la suite de ce mémoire, ce corpus Reuters-21578 sera simplement nommé *corpus Reuters.*

```
SHEARSON LEHMAN NAMES NEW MANAGING DIRECTOR
Shearson Lehman Brothers, a unit of
American Express Co &lt;AXP>, said Robert Stearns has joined the
company as managing director of its merger and acquisition
department.
    Shearson said Stearns formerly was part of Merrill Lynch
Pierce, Fenner and Smith Inc's &lt;MER> merger and acquisitions
department.
```

Figure 3.1 : *Exemple de texte du corpus Reuters-21578.*

Le corpus Reuters est souvent utilisé lors d'évaluation dans les publications, comme dans [Schapire *et al.*, 1998] pour comparer leur algorithme *AdaBoost* avec la formule de Rocchio, ou dans [Joachims, 1998] et [Dumais *et al.*, 1998] pour évaluer les performances des machines à vecteurs supports. [Yang et Liu, 1999] ont également utilisé ce corpus pour comparer différents algorithmes (machines à vecteurs supports, réseaux de neurones, arbres de décision, réseaux bayesiens).

Il est possible, sur ce corpus, de comparer les performances de l'approche proposée dans ce mémoire avec celles d'autres approches. Cependant, pour pouvoir faire ces comparaisons, il est nécessaire de remplir plusieurs conditions : d'une part, la base de test servant à l'évaluation des performances doit être identique pour toutes les méthodes, et d'autre part, les performances doivent être évaluées avec les mêmes mesures.

Afin de faciliter les comparaisons ultérieures, nous présentons, dans ce paragraphe, le découpage que nous avons considéré dans toute la suite de ce mémoire.

3.1.1 Distinction entre bases d'apprentissage et de test : le découpage Apté

Le découpage le plus souvent rencontré se nomme découpage *Apté* du nom des premiers auteurs à l'avoir proposé [Apté *et al.*, 1994]. La base d'apprentissage est constituée des documents antérieurs au 8 avril 1987, soit 9603 documents, et la base de test de tous les documents ultérieurs, soit 3299 documents.

Les catégories retenues sont celles pour lesquelles il existe au moins un document pertinent sur la base d'apprentissage et un document pertinent sur la base de test, ce qui permet de retenir 90 catégories différentes. Certains documents de la base de test peuvent appartenir à plusieurs catégories, d'autres à aucune.

C'est ce découpage que nous avons utilisé dans tout notre travail.

Comme nous l'avons déjà souligné au chapitre précédent, il existe malheureusement de légères modifications à ce découpage qui rendent certaines comparaisons difficiles. Ainsi [Yang et Liu, 1999] ont supprimé de la base de test tous les documents qui n'appartiennent à aucune catégorie : ils n'utilisent que 3019 documents sur la base de test. La suppression de ces documents ne peut qu'améliorer les résultats par rapport au découpage traditionnel, puisque les risques de mauvais classement sont réduits. [Dumais *et al.*, 1998] considèrent 118 catégories : certaines catégories n'ont donc pas de documents pertinents sur la base de test, et la façon dont ces catégories sont prises en considération dans leur évaluation n'est pas très claire.

3.1.2 Définition des catégories du corpus Reuters-21578

Les 90 catégories issues du découpage sont présentées à la Figure 3.2, ainsi que le nombre de documents pertinents disponibles sur chaque base. Elles sont classées par ordre décroissant du nombre de documents pertinents sur la base d'apprentissage. Le nombre de documents pertinents disponibles pour effectuer l'apprentissage décroît rapidement ; dès la vingt-sixième catégorie, ce nombre est inférieur à cinquante. Il faut noter que les documents pertinents sont à peu près également répartis sur les deux bases, c'est-à-dire que les catégories ayant beaucoup (respectivement peu) de documents pertinents sur la base d'apprentissage ont également beaucoup (respectivement peu) de documents pertinents sur la base de test.

La difficulté de ces catégories est variable : [Wiener *et al.*, 1995] ont montré que les distinctions entre certaines catégories reposent presque exclusivement sur la présence ou l'absence d'un mot-clef, (cette étude repose sur le corpus Reuters-22173, mais les conclusions sont vraies pour la nouvelle version du corpus).

	Catégorie	Apprentissage	Test
1	earn	2877	1087
2	acq	1650	719
3	money-fx	538	179
4	grain	433	149
5	crude	389	189
6	trade	369	118
7	interest	347	131
8	wheat	212	71
9	ship	197	89
10	corn	182	56
11	money-supply	140	34
12	dlr	131	44
13	sugar	126	36
14	oilseed	124	47
15	coffee	111	28
16	gnp	101	35
17	gold	94	30
18	veg-oil	87	37
19	soybean	78	33
20	nat-gas	75	30
21	bop	75	30
22	livestock	75	24
23	cpi	69	28
24	reserves	55	18
25	cocoa	55	18
26	carcass	50	18
27	copper	47	18
28	jobs	46	21
29	yen	45	14
30	ipi	41	12
31	iron-steel	40	14
32	cotton	39	20
33	gas	37	17
34	barley	37	14
35	rubber	37	12
36	alum	35	23
37	rice	35	24
38	palm-oil	30	10
39	meal-feed	30	19
40	sorghum	24	10
41	retail	23	2
42	zinc	21	13
43	silver	21	8
44	pet-chem	20	12
45	wpi	19	10

	Catégorie	Apprentissage	Test
46	tin	18	12
47	rapeseed	18	9
48	orange	16	11
49	housing	16	4
50	strategic-metal	16	11
51	hog	16	6
52	lead	15	14
53	soy-oil	14	11
54	heat	14	5
55	soy-meal	13	13
56	fuel	13	10
57	lei	12	3
58	sunseed	11	5
59	dmk	10	4
60	lumber	10	6
61	tea	9	4
62	income	9	7
63	oat	8	6
64	nickel	8	1
65	l-cattle	6	2
66	groundnut	5	4
67	instal-debt	5	1
68	rape-oil	5	3
69	platinum	5	7
70	sun-oil	5	2
71	jet	4	1
72	coconut	4	2
73	coconut-oil	4	3
74	potato	3	3
75	propane	3	3
76	cpu	3	1
77	copra-cake	2	1
78	palmkernel	2	1
79	naphtha	2	4
80	palladium	2	1
81	rand	2	1
82	dfl	2	1
83	nzdlr	2	2
84	rye	1	1
85	cotton-oil	1	2
86	lin-oil	1	1
87	castor-oil	1	1
88	sun-meal	1	1
89	groundnut-oil	1	1
90	nkr	1	2

Figure 3.2 : *Définition des catégories, avec le nombre de documents pertinents disponibles pour chaque partie de la base.*

3.2 Le corpus TREC-8

La conférence TREC (TREC)[1] a été présentée brièvement au premier chapitre avec les trois sous-tâches de relatives au filtrage : le filtrage adaptatif, le filtrage par lots (*batch*) et le routage (*routing*)[2].

Parmi ces trois sous-tâches, le filtrage adaptatif est considéré comme le moins bien adapté à la mise en œuvre de méthodes statistiques d'apprentissage, et le routing comme le mieux adapté à l'application de ces méthodes. Nous verrons, en effet, dans le paragraphe suivant, que le routing est la sous-tâche pour laquelle le nombre de documents pertinents est le plus élevé.

3.2.1 Description des données utilisées pour le filtrage dans TREC-8

Pour la compétition TREC-8, il fallait construire cinquante profils correspondant à cinquante requêtes différentes, identifiées par un numéro allant de 351 à 400. La Figure 3.3 présente deux requêtes proposées pour la compétition. Elles sont définies par un titre, une description qui détaille le titre et une partie narrative qui précise exactement ce que doivent être les documents pertinents, et également ce qu'ils ne doivent pas être. Ces deux exemples montrent que les requêtes sont assez précises, donc difficiles à satisfaire.

À l'époque de la compétition, les documents pertinents pour chacune des requêtes provenaient de plusieurs corpus : le *Financial Times* 1992, 1993 et 1994 (noté FT92, FT93, FT94), *Federal Registe*r 1994 (FR94), *Congressional Record* 1993 (CR), *Foreign Broadcast Information Service* (FBIS), et *LA Times*.

[1] http://trec.nist.gov

[2] Dans la suite de ce mémoire, nous utilisons indifféremment les mots *routing* et *routage.*

```
<num> Number: 351
<title> Falkland petroleum exploration

<desc> Description:
What information is available on petroleum exploration in
the South Atlantic near the Falkland Islands?

<narr> Narrative:
Any document discussing petroleum exploration in the
South Atlantic near the Falkland Islands is considered
relevant.  Documents discussing petroleum exploration in
continental South America are not relevant.


<num> Number: 352
<title> British Chunnel impact

<desc> Description:
What impact has the Chunnel had on the British economy and/or
the life style of the British?

<narr> Narrative:
Documents discussing the following issues are relevant:

 - projected and actual impact on the life styles of the British
 - Long term changes to economic policy and relations
 - major changes to other transportation systems linked with
   the Continent

Documents discussing the following issues are not relevant:

 - expense and construction schedule
 - routine marketing ploys by other channel crossers (i.e.,
   schedule changes, price drops, etc.)
```

Figure 3.3 : *Deux exemples de requêtes pour la compétition TREC-8.*

La Figure 3.4, qui est un exemple de texte extrait du corpus *Financial Times*, montre que, comme sur le corpus Reuters, les textes sont rédigés.

```
FT 21 SEP 93
Hounslow to cut 250 jobs
HOUNSLOW, the Labour-controlled London borough, is to shed 250 jobs as part
of a Pounds 6m cuts package it said would be necessary to avoid budget-
capping by central government.
It is the first council to announce its planned budget for next year, and
it said the cuts had been made necessary by the government's plan to cut
budgets by 2 per cent.
Hounslow's cut is equivalent to 4.7 per cent of its controllable budget and
the council predicted that other authorities would have to follow suit. Mr
John Chatt, the council leader, said: 'We have gone as far as we can in
recommending cuts which will minimise the effect on direct services to the
community. If we don't get a decent grant settlement then any further cuts,
and the damage they will cause, will be due to the government's failure to
listen to our pleas.'
(…)
```

Figure 3.4 : *Exemple de texte issu du Financial Times.*

La Figure 3.5 précise, pour chacune des différentes sous-tâches, les bases à utiliser ; comme indiqué plus haut, le routing bénéficie du plus grand nombre de documents pertinents pour la base d'apprentissage.

	Adaptatif	Filtrage par lots	Routage
Apprentissage	-	FT92	FT92, FR94, LA CR, FBIS
Test	FT92, FT93, FT94	FT93, FT94	FT93, FT94

Figure 3.5 : *Répartition des exemples en base d'apprentissage et base de test selon les différentes sous tâches.*

Pour le routing, la Figure 3.6 indique le nombre de documents pertinents disponibles pour chaque thème, ainsi que la répartition entre les bases ; le nombre moyen de documents pertinents pour la base d'apprentissage est de 66, et la médiane se situe à 55.

La partie *Financial Times* 92 du corpus comporte 64139 documents. La base de test (FT93, FT94) est composée de 140650 documents ; le nombre moyen de documents pertinents sur la base de test est de 46,7 documents par thème, soit une proportion moyenne de 0,03 %.

Thème	Total	FT92	FBIS	FR	LA	Test (FT93-94)
351	31	11	19	0	1	17
352	62	55	3	0	4	190
353	88	15	34	8	31	29
354	311	21	175	0	115	49
355	44	1	38	2	3	1
356	10	8	0	1	1	15
357	230	24	205	0	1	61
358	51	0	0	0	51	2
359	10	4	0	0	6	30
360	139	2	89	0	48	12
361	8	0	1	0	7	2
362	36	0	17	0	19	5
363	12	2	1	0	9	4
364	33	1	16	2	14	2
365	32	8	1	7	16	3
366	84	4	17	20	43	17
367	164	12	94	4	54	30
368	56	0	2	4	50	5
369	13	1	0	1	11	0
370	326	15	0	267	44	31
371	16	1	1	10	4	1
372	37	3	0	17	17	12
373	23	2	0	17	4	12
374	150	19	22	0	109	59
375	76	9	34	1	32	10
376	87	13	65	0	9	17
377	29	5	5	3	16	10
378	61	51	9	0	1	44
379	16	0	0	0	16	0
380	6	0	1	0	5	1
381	26	5	8	2	11	2
382	18	2	10	2	4	4
383	75	15	0	5	55	67
384	48	1	34	0	13	3
385	61	9	30	1	21	27
386	16	1	0	3	12	6
387	76	4	57	12	3	9
388	39	3	6	1	29	14
389	95	41	32	2	20	129
390	85	2	1	12	70	68
391	67	55	0	0	12	127
392	82	27	28	1	26	32
393	66	4	3	0	59	5
394	12	0	3	1	8	5
395	156	36	102	1	17	62
396	55	6	1	27	21	5
397	21	2	0	0	19	5
398	136	10	66	0	60	10
399	96	4	60	14	18	9
400	109	34	49	0	26	16

Figure 3.6 : *Nombre de textes pertinents pour chaque thème.*

3.2.2 Fabrication du fichier des pertinences avant la conférence TREC-8

La constitution d'une base de documents étiquetés de grande taille est un problème complexe. Nous exposons ici la méthode retenue par NIST pour constituer ces bases, car elle a des conséquences importantes.

L'ensemble des documents représente environ cinq Gigaoctets de données ; il n'est donc évidemment pas possible de lire chaque document pour vérifier sa pertinence par rapport à l'ensemble des thèmes.

Le fichier des pertinences pour TREC-8 a été fabriqué grâce à la tâche *ad hoc* de TREC-7. Pour cette recherche *ad hoc*, les participants devaient fournir, pour chacune des 50 requêtes 351 à 400, les 1000 documents les plus pertinents, en ne disposant que des requêtes. Ensuite, pour chacune des requêtes, les cent premiers documents fournis par chaque candidat ont été assemblés pour former un ensemble soumis à des assesseurs qui classent chacun des documents de l'ensemble : pertinent ou non pertinent pour la requête. La liste des documents soumis aux assesseurs est ordonnée par ordre chronologique : ils ne savent pas à quel rang était classé chaque document, ni par quel système il a été sélectionné.

Grâce à ce travail, les bases sont étiquetées pour les requêtes 351 à 400, qui peuvent alors être utilisées pour TREC-8. L'ensemble des documents est ainsi divisé en trois parties : les documents étiquetés par les assesseurs comme pertinents, les documents étiquetés par les assesseurs comme non pertinents, et les documents qui ne faisaient pas partie de l'ensemble à étiqueter et qui sont considérés comme non pertinents.

On peut donc considérer qu'il existe deux sous-ensembles de documents non pertinents :

- les documents qui avaient été sélectionnés par un système et dont la non-pertinence a été vérifiée par un assesseur.
- les documents qui n'ont été sélectionnés par aucun système et qui n'ont donc jamais été lus par les assesseurs ; ils sont considérés comme non pertinents.

À l'aide de ce fichier, disponible pour chaque requête, il est possible, grâce aux documents reconnus comme pertinents, d'effectuer des apprentissages pour la sous-tâche du routing ou du filtrage par lots.

3.2.3 Fabrication du fichier des pertinences après la conférence TREC-8

Pour TREC-8, il avait été décidé qu'une nouvelle étude de la pertinence des documents relativement aux requêtes 351 à 400 serait effectuée grâce aux résultats fournis par les candidats de TREC-8, selon le principe expliqué au paragraphe précédent.

Grâce à cette nouvelle évaluation, de nouveaux documents pertinents ont été trouvés pour plusieurs requêtes sur l'ensemble de la base FT92, FT93 et FT94 ; les performances officielles de la compétition ont été calculées avec ces nouveaux fichiers de documents pertinents.

3.2.4 Conséquence importante

Comme il y a eu deux fichiers successifs de documents pertinents, il faut savoir exactement quels fichiers sont utilisés. Pour comparer les résultats d'une étude aux résultats de la compétition, il faut utiliser le fichier des documents pertinents disponible *avant* la compétition pour effectuer les apprentissages : sinon, on dispose de plus de documents pertinents pour l'apprentissage qu'avant la compétition. Il faut ensuite évaluer les résultats en utilisant le fichier des documents pertinents délivré *après* la compétition pour la base de test FT93, FT94 puisque les performances des systèmes à la compétition ont été évaluées avec ce fichier.

Il faut noter qu'il existe alors un léger biais, puisque les nouveaux documents proposés par une nouvelle méthode ne sont pas évalués par des assesseurs comme lors de la compétition.

3.3 Conclusion

Le corpus Reuters présente l'avantage d'être de taille modérée, et de représenter une grande variété de situations, avec des catégories comportant beaucoup d'exemples pertinents, et d'autres qui en comportent très peu.

Une des particularités du corpus TREC-8 pour la tâche de routing est que les exemples pertinents sont issus de corpus différents ; nous n'avons pas étudié l'impact exact de cette hétérogénéité, mais il semble que, pour certaines catégories, cela soit préjudiciable.

Il faut noter que ces deux corpus sont composés de textes journalistiques issus d'organes de presse de qualité puisqu'il s'agit de l'agence Reuters et du Financial Times. Ces textes sont rédigés et structurés dans une langue correcte et sont proches des dépêches de l'AFP traitées au chapitre 9. Si l'on avait travaillé sur la classification des courriers électroniques, il aurait été préférable de choisir d'autres corpus.

Il existe cependant une différence importante entre ces corpus et les dépêches AFP : ils sont en langue anglaise alors que les dépêches AFP sont en langue française. Mais toutes les méthodes utilisées reposent sur les statistiques d'apparition des mots et sont donc à peu près indépendantes de la langue, tant que les mots sont définis comme des chaînes de caractère entourées d'espace ou de signes de ponctuations (ce qui n'est pas nécessairement le cas pour toutes les langues, notamment certaines langues asiatiques).

Néanmoins, des exemples de thèmes avec les dépêches AFP sont également utilisés dans ce mémoire, pour vérifier que le comportement des méthodes n'est pas modifié avec la langue française. Il en effet facile d'utiliser l'application ExoWeb du chapitre 9 pour disposer d'un ensemble important de documents étiquetés comme pertinents puisqu'il existe déjà un système de filtrage à la Caisse des Dépôts.