ANNEXE A : Présentation du corpus TREC-9

A.1 Présentation du corpus pour la tâche de filtrage de TREC-9

Ce paragraphe présente le corpus utilisé pour la tâche de filtrage de TREC-9.

Ce corpus est composé de la collection OHSUMED qui est un ensemble extrait da la base MEDLINE. Il est composé d'extraits de journaux médicaux datés de 1987 à 1991. En général, ces textes comportent un titre et un résumé, mais certains d'entre eux peuvent ne comporter qu'un titre. Ces textes contiennent, en plus, des annotations manuelles qui correspondent à des catégories manuelles appelées *Medical Subject Headings* ou *MeSH*.

Tous les textes du corpus sont composés selon le principe de la Figure A.1 : la référence de la revue, les annotations manuelles, le titre de l'article, le type de la revue, le résumé (s'il existe) et les noms des auteurs.

```
.S
Référence de la revue
.M
Annotations manuelles
.T
Titre de l'article
.P
Type de la revue
.W
Résumé
.A
Auteurs
```

Figure A.1 : *Composition des textes du corpus TREC-9.*

La Figure A.2 est un exemple de texte de ce corpus avec la composition précédente.

```
.S
Dig Dis Sci 8705; 31(12):1313-6
.M
Adult; Age Factors; Aged; Animal; Breath Tests; Colonic Diseases,
Functional/DH/ET; Female; Food Habits; Human; Italy; Lactose; Lactose
Intolerance/DH/DI/*EP; Male; Middle Age; Milk; Sex Factors.
.T
Lactose malabsorption and intolerance in Italians. Clinical implications.
.P
JOURNAL ARTICLE.
.W
Lactose malabsorption was assessed by the hydrogen breath test in 40
Italian patients with irritable bowel syndrome and 42 controls without
abdominal disturbances. Sixty-five percent of patients were "low milk
consumers" (0-250 ml milk per day) compared with 38% of controls (P less
than 0.02). Lactose loads of 25 and 50 g caused malabsorption in 82.5 and
87.5% patients and in 55 and 62% controls, respectively (patients vs
controls P less than 0.02). Malabsorption was more frequent in the "low
milk consumers" group (P less than 0.05).
.A
```

```
Bozzani A; Penagini R; Velio P; Camboni G; Corbellini A; Quatrini M.
```

Figure A.2 : *Exemple de texte du corpus TREC-9.*

La partie du corpus datée de 1987 est utilisé pour fabriquer les bases d'apprentissages et contient 54710 documents différents. La base de test est constituée des 293882 documents datés entre 1988 et 1991.

Ce corpus a donc une taille plus grande que celui utilisé pour TREC-8 où la base de test comportait 140650 documents.

A.2 Définition des thèmes et des documents pertinents

Pour la sous-tâche de routing de TREC-9, trois ensemble de thèmes étaient proposés :

1. 63 thèmes appelés *OHSUMED queries*.
2. 4904 thèmes appelés *MeSH queries* construits à partir des annotations manuelles.
3. 500 thèmes choisis parmi les 4904 thèmes précédents appelés *MeSH sample.*

Le troisième ensemble a été constitué, car il est apparu, au travers des discussions entre les participants de TREC, que le traitement des 4904 thèmes nécessitait un temps de calcul considérable compte tenu de la taille importante de la base de test. L'ensemble des 500 thèmes devait permettre à un plus grand nombre de proposer des résultats pour ces thèmes.

En pratique, aucun candidat n'a utilisé l'ensemble des 4904 thèmes ; nous n'en reparlons plus dans la suite.

La figure A.3 montre deux exemple de requêtes qui, comme pour TREC-8, comporte un titre et une partie narrative.

```
<top>
<num> Number: OHSU7
<title> young wf with lactase deficiency
<desc> Description:
lactase deficiency therapy options
</top>

<top>
<num> Number: MSH-SMP1
<title> Abdomen, Acute
<desc> Description:
Clinical syndrome characterized by abdominal pain of great severity
associated with other symptoms and signs, usually those of acute
peritonitis, which might well be the result of a ruptured abdominal viscus
or a similar abdominal catastrophe requiring urgent surgical operation.
```

```
</top>
```

Figure A.3 : *Deux exemples de requêtes : la première est issue de l'ensemble appelé OHSUMED queries et la deuxième de l'ensemble MeSH sample.*

La figure A.4 précise le nombre moyen de documents pertinents disponibles pour construire la base d'apprentissage sur chaque ensemble de thèmes. Ce nombre est très faible pour les requêtes OHSUMED ; il est beaucoup plus faible que pour la tâche de routing de TREC-8.

	OHSUMED queries	*MeSH Sample*
Nombre de requêtes	63	500
Nombre moyen de documents pertinents sur la base d'apprentissage	10,6	46,5
Médiane	8	25

Figure A.4 : *Nombre de documents pertinents pour la base d'apprentissage.*

A.2 Utilisation des annotations manuelles

Les directives de TREC-9 précisaient clairement comment devait être utilisé les annotations manuelles figurant dans les textes. Pour l'ensemble des thèmes OHSUMED, il était possible de les utiliser à condition de le préciser, mais pour les thèmes *MeSH sample*, leur utilisation était interdite, car ces thèmes ont été construits à partir de ces annotations.

En pratique, l'utilisation des annotations manuelles consiste à considérer le texte qui suit le champ ".M" comme faisant partie du texte avec le titre et le résumé.

A.3 Notre participation à la sous-tâche de routing de TREC-9

Il est possible d'envoyer plusieurs fichiers résultats appelés *run* pour l'évaluation.

Nous avons soumis trois *run* pour la tâche de routing dont les noms et les caractéristiques sont présentés à la Figure A.5.

	Ensemble des requêtes	Utilisation des annotations manuelles
S2RNr1	*OHSUMED queries*	non
S2RNr2	*OHSUMED queries*	oui
S2RNsamp	*MeSH Sample*	non

Figure A.5 : *Description des résultats envoyés à TREC-9.*

ANNEXE B :

Article :

Vers la conception automatique de filtres d'informations efficaces.

Reconnaissance des formes et Intellignece Artificielle (RFIA 2000), 129-137, 2000.

ANNEXE C :

Articles :

Using Learning-Based Filters to Detect Rule-Based Filtering Obsolescence.

Proceedings of the Sixth Conference on Content-Based Multimedia Information (RIAO 2000), 1208-1220, 2000.

ANNEXE D :

Article :

Two Step Feature Selection for the TREC-8 Routing.

Proceedings of the Eighth Text REtrieval Conference (TREC-8).

NIST Special Publication 500-246, 425-430, 2000.

ANNEXE E :

Article :

Training Context-Sensitive Neural Networks With Few Relevant Examples for the TREC-9 Routing.

Proceedings of the Ninth Text REtrieval Conference (TREC-9).

NIST Special Publication, A paraître, 2001.