Comment trouver "LA" réponse

Olivier Ferret, Brigitte Grau, M Huraut-Plantet, Gabriel Illouz, Christian Jacquemin

Comment trouver “ LA ” réponse ?

Olivier Ferret, Brigitte Grau, Martine Hurault-Plantet, Gabriel Illouz et Christian Jacquemin

LIMSI-CNRS, BP 133 - 91403 Orsay cedex

[ferret,grau,mhp,illouz,jacquemin]@limsi.fr

Résumé :

Nous avons développé le système QALC permettant de trouver la réponse à des questions précises dans un grand corpus. Cette problématique est à l'intersection de deux domaines : la recherche d'information et le traitement automatique des langues. Ainsi QALC réalise une analyse de documents sélectionnés par un moteur de recherche en se fondant sur la reconnaissance des termes de la question et de leurs variantes, ainsi que sur la reconnaissance des entités nommées qu'ils contiennent. Ces éléments donnent des indices pour sélectionner un nombre minimum de documents et comparer chacune de leurs phrases avec la question traitée.

Mots Clés : Recherche d'information, Traitement automatique des langues, Systèmes de question/réponse.

Abstract :

We developed a system, QALC, in order to find answer to precise questions in a very large corpus. This kind of problem refers to two domains: information research and natural language processing. Thus QALC realizes an analysis of documents selected by a search engine based on the search for multi-word terms and their variations, and the recognition of their named entities. These indexes are used to select a minimal number of documents to be processed and to give indices when comparing question and sentence representations.

Keywords: Information retrieval, Natural language processing, Question Answering.

1. Introduction

Les systèmes de recherche d'information, pour être vraiment utilisables, doivent répondre à des besoins précis en matière d'information. Lorsque l'intérêt de l'utilisateur porte sur des thèmes particuliers ou sur le suivi de l'évolution d'un domaine, il est pertinent d'utiliser des systèmes de filtrage et de résumé. Lorsque ce même utilisateur recherche une donnée factuelle, l'information pertinente ne pourra être apportée que par des systèmes dédiés à ce type de tâche. En effet, face à une question telle que " Quelle est la voiture la plus chère du monde ? ", les moteurs de recherche traditionnels renvoient tous les documents où figurent les mots de la question et c'est à l'utilisateur que revient la tâche d'explorer ces documents afin de trouver la réponse. Répondre à des questions précises requiert une analyse plus en profondeur des documents afin d'en extraire l'information pertinente. À cette fin, nous avons développé le système QALC [Ferret et al, 2000] qui allie des techniques issues de la recherche d'information et du traitement automatique des langues (TAL).

2. Architecture du système QALC

Un point important dans un système de question/réponse est de restreindre le plus possible le nombre de documents dans lesquels s'effectue la recherche de la réponse. À cette fin, nous retenons un ensemble assez large de documents fournis par un moteur de recherche (cf. figure 1), ensemble sur lequel s'opère une sélection reposant sur la reconnaissance des termes de la question ou de leurs variantes morphologiques, syntaxiques ou sémantiques et non pas seulement sur des mots simples des textes. Cette reconnaissance est effectuée par FASTR [Jacquemin, 1999] sur la base des termes extraits de la question grâce à des expressions régulières décrivant des groupes nominaux complexes et leurs sous-parties.

La recherche d'une réponse précise dans un document est guidée par un certain nombre d'indices qui s'ajoutent à l'utilisation des mots de la question. L'un de ces indices est le type de réponse attendu, en terme de type d'entités nommées que le système sait reconnaître, c'est-à-dire des expressions désignant des personnes, des organisations, des lieux et des quantités. La reconnaissance de ces entités dans les documents est fondée à la fois sur des règles lexico-syntaxiques et sur la consultation de lexiques, généraux et spécialisés. Un autre indice concerne la prise en compte de la variation linguistique. QALC utilise donc les résultats de FASTR lors de la dernière phase de recherche de la réponse.

Ce dernier module, qui propose un ensemble limité de réponses à chaque question, met en œuvre un calcul de similarité entre une question et une phrase d'un document sélectionné sur la base de leurs mots pleins lemmatisés, de leurs termes et du type des entités nommées, attendues dans le cas de la question, reconnues dans le cas de la phrase.

Les réponses données par QALC sont des phrases, unités les plus significatives du point de vue de l'utilisateur final, mais peuvent être rendues plus concises grâce à un ensemble d'heuristiques permettant d'extraire de la phrase une réponse précise.

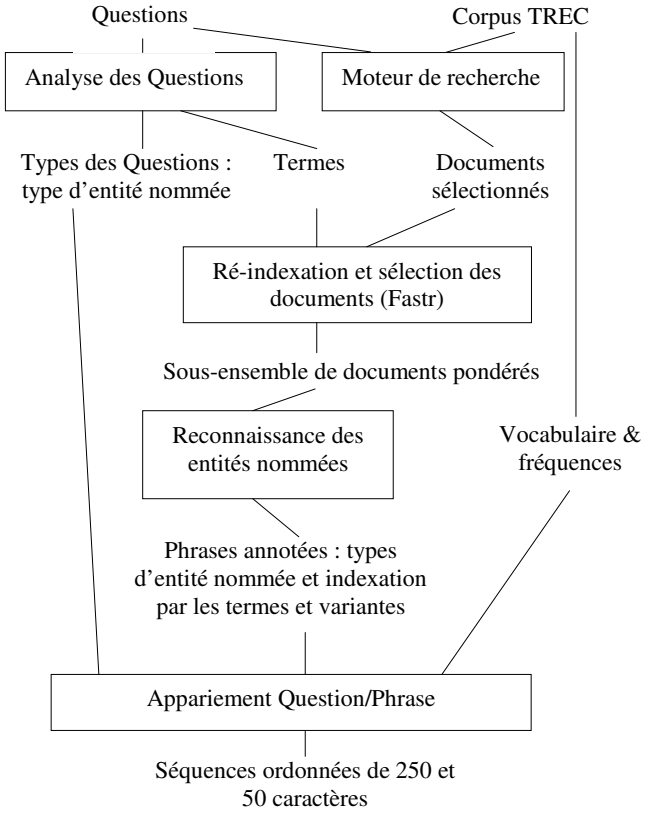


Figure 1: Architecture du système QALC

Nous allons présenter par la suite l'ensemble des modules de QALC.

3. Analyse des questions

L'analyse des questions est fondée sur l'application d'un ensemble de règles exploitant des critères lexicaux, syntaxiques et sémantiques. Les étiquettes sont hiérarchisées (cf. figure 2) afin d'offrir plus de souplesse lors de la comparaison d'une phrase et d'une question. La détermination du type attendu permet de le mettre en relation avec l'entité nommée correspondante dans les phrases candidates. Par exemple :

Question : *How much could you rent a Volkswagen bug for in 1966? —>* FINANCIER

(Pour combien pouvait-on louer une coccinelle Volkswagen en 1966 ?)

Réponse : *... When Dollar Rent-a-Car opened in 1966, you could rent a Volkswagen bug for* **<bnumex TYPE=FINANCIER>** *$1* **<enumex>** *a day.*

(Quand Dollar Rent-a-Car ouvrit en 1966, vous pouviez louer une coccinelle pour 1$ par jour.)

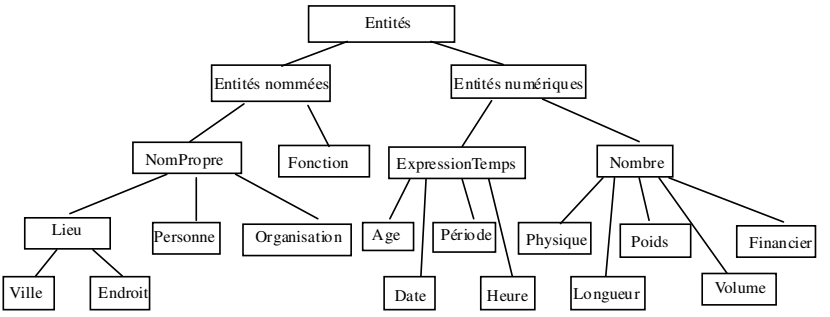


Figure 2: Hiérarchie des types de réponses

4. Ré-indexation et sélection des documents

4.1. Seuil de sélection des documents

Afin d'évaluer le nombre de documents à retenir parmi ceux qui sont renvoyés par le moteur de recherche, nous avons effectué différents tests en faisant varier le nombre de documents retenus, 50, 100, 200 et 500, en comptant les questions pour lesquelles ces documents contenaient la réponse cherchée.

Table 1. Nombre de questions avec et sans documents pertinents selon le seuil

Seuil de sélection	Questions *avec* des documents pertinents	Questions *sans* documents pertinents
50	181	19
100	184	16
200	193	7
500	194	6

On peut voir dans la table 1 que le nombre de questions susceptibles d'être résolues n'évolue pas au delà de 200 documents, seuil que nous avons retenu dans

QALC et qui semble être le meilleur compromis entre temps de calcul et perte d'information.

4.2. Ré-indexation par les termes et leurs variantes

L'indexation automatique des documents retenus en fonction des termes de la question est faite par FASTR, un analyseur transformationnel de surface pour la reconnaissance de variantes terminologiques. Les termes sont extraits de la question à l'aide d'un patron décrivant la composition d'un groupe nominal. Ils correspondent au groupe nominal reconnu dans son entier et à ses sous parties. Les patrons de variations des termes qui reposent sur des familles morphologiques et sémantiques sont engendrés au moyen de métarègles à partir de la description des termes reconnus.

La *famille morphologique* d'un mot simple *m* est l'ensemble *M*(*m*) des mots simples de la base CELEX [CELEX, 1998] qui ont la même racine que *m*. Par exemple, la famille morphologique du nom *maker* (fabricant) se compose des noms *maker*, *make* (marque) et *remake* (remake), et des verbes *to make* (faire) et *to remake* (refaire).

La *famille sémantique* d'un mot simple *m* est l'union *S*(*m*) des *synsets* de WordNet1.6 [Fellbaum, 1998] auxquels ce mot *m* appartient. Un synset est l'ensemble des mots qui partagent un lien de synonymie sur une de leurs entrées sémantiques. Par exemple, la famille sémantique de *maker* se compose de trois noms : *maker*, *manufacturer* (fabricant), *shaper* (façonneur) et la famille sémantique de *car* (voiture) est *car*, *auto*, *automobile*, *machine* et *motorcar* (voiture à moteur).

En s'appuyant sur les familles morphologiques et sémantiques et sur l'ensemble de métarègles pour l'anglais, les occurrences suivantes sont extraites comme des variantes du terme d'origine *car maker* (fabricant de voitures) :

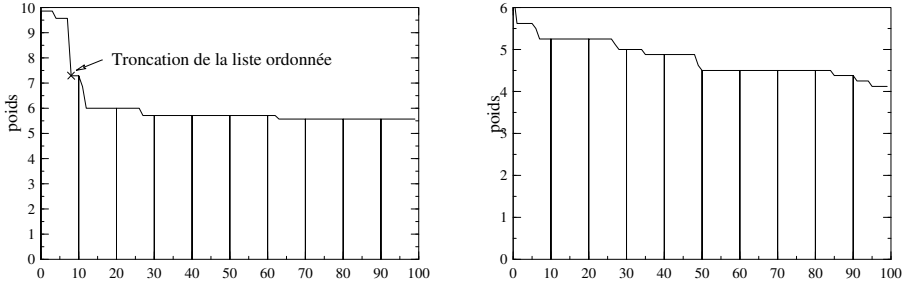
> *auto maker* (fabricant d'autos), *auto parts maker* (fabricant de pièces détachées automobiles), *car manufacturer* (fabricant de voitures), *make autos* (fabriquer des voitures) et *making many automobiles* (fabricant beaucoup de voitures).

4.3. Sélection des documents

FASTR produit la liste des variantes trouvées pour chaque document associé à une question. Ces variantes sont pondérées en fonction de la fiabilité de la variation trouvée. Le poids *w(v)* d'une variante *t(q,i)* correspondant exactement au terme de référence vaut 3 ; pour une variante morphologique ou morpho-syntaxique, il est égal à 2, et pour une variante sémantique ou morpho-sémantico-syntaxique, il vaut 1. Ce poids est renforcé en fonction de la proportion *P(t(q,i))* de noms propres dans le terme, du fait de leur fiabilité, et est modulé par la longueur du terme, $|t(q,i)|$: plus celui-ci comporte de mots, plus il est considéré comme fiable. Les pondérations de chacun des termes retenus en tant qu'index pour le document *d* sont sommés et normalisés en fonction du nombre de termes $|T(q)|$ dans la question *q*. Le poids du document *d* est donc donné par la formule suivante :

$$W_q(d) = \sum_{i \in I(d)} \frac{w(v) \times (1 + 2P(t(q,i))) \times |t(q,i)|}{|T(q)|} \qquad (1)$$

Ce poids est calculé pour les 200 documents retenus pour chaque question. La distribution de ces poids permet de réaliser un filtrage plus sélectif des documents. On observe principalement deux types de courbes de pondération des documents sélectionnés pour une question : les courbes avec un plateau et une chute brutale des valeurs des poids au-delà d'un certain rang (figure 3.a) et les courbes avec des valeurs de poids en décroissance progressive (figure 3.b).



(a) Question numéro 87 **(b)** Question numéro 86

Figure 3 : Deux types de courbes de pondération

Lorsque le système détecte une pente suffisamment forte, il sélectionne les documents apparaissant avant la chute ; sinon il retient un nombre de documents fixé *a priori* (nombre égal à 100). Dans le cas de la figure 3.a, qui correspond à la question 87 des données de TREC8 *Who followed Willy Brandt as chancellor of the Federal Republic of Germany?* (Qui a succédé à Willy Brandt comme chancelier de la République Fédérale d'Allemagne ?), QALC a trouvé un seuil égal à 8 documents. En revanche la courbe de la figure 3.b ne présente pas de seuil détectable. Pour la question 86 correspondante *Who won two gold medals in skiing in the Olympic Games in Calgary?* (Qui a gagné deux médailles d'or en ski aux Jeux Olympiques de Calgary ?), QALC a donc retenu 100 documents.

4.4. Évaluation de la sélection des documents

Nous avons évalué l'efficacité du filtrage en appliquant notre chaîne de traitement sur les données de TREC8, une fois avec le processus de filtrage, et une autre sans cette sélection. Sans filtrage, les 200 documents issus du moteur de recherche étaient donc conservés pour chacune des 200 questions traitées. Nos tests ont donné un score de 0,463 dans le premier cas, et de 0,452 dans le second. Ces résultats montrent que les performances ne diminuent pas en traitant moins de documents car la ré-indexation conduit à classer en premier les documents les plus pertinents. D'ailleurs, les performances de notre système sont généralement

meilleures lorsque le filtrage est opérant et conserve moins de 100 documents, comme le montre la table 2.

Table 2. Évaluation de la sélection

Nombre de documents sélectionnés	**100**	**<<100**
Distribution parmi les questions	342 (50%)	340 (50%)
Nombre de réponses correctes	175 (51%)	200 (59%)
Nombre de réponses correctes au rang 1	88 (50%)	128 (64%)

5. Sélection de la réponse

Le module de sélection de la réponse considère la phrase en tant qu'unité de réponse de base dans la mesure où elle apparaît comme le meilleur compromis entre une nécessaire concision et la conservation d'un contexte permettant de juger la pertinence de la réponse. La sélection des réponses possibles s'effectue de ce fait par appariement entre la question et chacune des phrases des documents conservés à l'issue de la phase de filtrage. Plus précisément, le module de QALC en charge de cet appariement question/phrase gère en permanence une liste réduite des N_r[1] phrases les plus à même de contenir la réponse à la question posée.

L'appariement entre une phrase et une question évalue la pertinence de cette phrase en fonction des entités caractéristiques de la question au sein la phrase. Ces entités sont des mots pleins, des termes ou des types d'entités nommées. Chacune d'elles est pondérée en fonction de son importance relative vis-à-vis des autres entités du même type et de son degré de ressemblance par rapport à l'entité homologue de la question. Un score est calculé pour chaque type d'entité à partir de ces poids et rend compte de la proximité de la phrase avec la question selon le point de vue représenté par ce type d'entités. Ces scores sont ensuite combinés afin de rendre un avis global sur la similarité de la phrase par rapport à la question. Les N_r phrases conservées à un moment donné sont celles ayant le score global le plus élevé parmi celles déjà analysées.

Les mots pleins résultent d'un étiquetage morpho-syntaxique et d'une lemmatisation réalisés par l'outil *TreeTagger* [Schmid, 1999]. Seuls les mots pleins de la phrase présents dans la question sont retenus et se voient attribuer un poids, de type *tf.idf*, en rapport avec leur fréquence et leur répartition dans le corpus source des réponses. Au niveau de la phrase, les termes considérés sont les variantes des termes de la question, reconnues par FASTR, avec leur poids. Enfin, les types d'entités nommées attendus pour la question sont mis en correspondance avec les types d'entités nommées qui ont été identifiées dans la phrase. Ces types formant une hiérarchie (cf. figure 2), le poids d'un type dans la phrase est d'autant plus faible qu'il est un sur-type éloigné du type attendu de la réponse.

[1] N_r est égal à cinq phrases dans le cas de l'évaluation *Question Answering* de TREC.

Le score caractérisant chaque type d'entité dans une phrase P, relativement à une question Q, est égal à la somme des poids des entités de ce type dans P. Le score global de P combine ces trois scores au travers de la formule suivante :

$$\begin{aligned} \text{score}(P/Q) = \text{score\_motsPleins}(P/Q) \cdot \alpha + \text{score\_termes}(P/Q) \cdot \beta + \\ \text{score\_entitésNommées}(P/Q) \cdot \gamma \end{aligned} \quad (2)$$

où α, β et γ sont des poids permettant d'ajuster l'importance relative accordée à ces trois types d'entités.

Dans le cas où l'on souhaite des réponses d'une taille inférieure à la phrase, QALC s'appuie sur un ensemble d'heuristiques simples pour réduire la taille des phrases dépassant la limite fixée. Lorsqu'une entité nommée correspondant au type attendu de la réponse, ou à un type proche, a été trouvée, QALC sélectionne la partie de la phrase entourant cette entité nommée. Dans le cas contraire, ou si le type attendu de la réponse n'a pu être déterminé, il extrait une partie de la phrase contiguë à celle contenant les mots de la question. On suppose ainsi que la phrase possède une structure comparable à ce que serait la forme affirmative de la question.

6. Résultats et discussion

L'évaluation des systèmes dans le cadre de la tâche *Question Answering* (QA) de TREC9 a consisté à proposer 5 réponses ordonnées, de 250 ou 50 caractères, pour chacune des 682 questions posées, réponses à retrouver parmi près d'un million de documents, principalement des articles de journaux américains. Un score est ensuite calculé faisant la moyenne de l'inverse des rangs des réponses correctes (0 si aucune réponse n'est correcte). Les résultats de QALC ont été évalués dans trois conditions différentes. Les variations concernent soit le moteur de recherche utilisé, avec les résultats du moteur ATT fournis par les organisateurs ou les résultats d'Indexal [Loupy et al., 1998], moteur fourni par Bertin Technologie, soit la taille de la réponse (250 ou 50 caractères). Le meilleur de ces tests a obtenu un score de 0,407 avec 375 réponses trouvées sur 682, pour des réponses de 250 caractères (voir table 3). Ce score nous a placé en 6$^{\text{ème}}$ position sur 28 participants.

Table 3. Nombre de réponses correctes trouvées pour les 2 tests à 250 caractères

Rang des réponses correctes	Test avec ATT	Test avec Indexal
1	216	187
2 à 5	159	185
Nombre de réponses correctes	375	372
Questions sans réponses	307	310

Un point important pour évaluer la similarité entre une question et une réponse porte sur le choix des termes de la question à retenir, choix d'autant plus crucial que

la question comporte peu de mots. Par exemple, pour la question *How far away is the moon ?*, notre module d'extraction des termes a conservé non seulement *moon (nom)*, mais aussi *away (adverbe)*, comme mots pleins pour l'appariement. D'autre part, notre module d'analyse de la question dispose, comme beaucoup d'autres, de *how far* comme locution interrogative permettant de typer la réponse attendue comme une distance, ce qui nous a permis de trouver la bonne réponse[2]. L'importance relative accordée aux différents termes de la question est un autre point délicat. Lorsque la question comporte un nom propre, il est important de retrouver ce nom propre dans la réponse ; il faut alors donner à ce terme un poids plus élevé. Ces considérations montrent que l'établissement d'une mesure de similarité faisant intervenir des entités de natures différentes, contrairement aux habitudes de la recherche d'information, est chose difficile, un type de terme important pour certaines questions pouvant se révéler source de bruit pour d'autres.

7. Travaux connexes

Il y a deux ans, la conférence TREC introduisait la tâche *Question Answering*. Cette conférence fournit un terrain d'expérimentation intéressant pour comparer l'efficacité des méthodes proposées. Même si les composants de base des systèmes sont analogues, les méthodes se révèlent plutôt différentes. Mais il est difficile de savoir précisément ce qui, dans des systèmes complexes de cette nature, apporte une amélioration ou non.

Comme dans QALC, la plupart des systèmes déterminent le type de réponse attendu par l'application de patrons prédéfinis. Prager et al. [Prager, 2000], par exemple, possèdent 400 modèles différents pour identifier environ 50 types de réponse. Le système développé par Ittycherian et al [Ittycherian, 2000] fonde la classification de ses types de réponse sur le modèle de l'entropie maximum. Pour leur part, les auteurs du système FALCON [Harabagiu, 2000] ont établi leur taxonomie à partir de la hiérarchie de WordNet.

Tous les systèmes participant à QA utilisent un moteur de recherche qui sélectionne un premier sous-ensemble de documents. Dans QALC, nous gardons les documents sélectionnés dans leur ensemble, alors que d'autres systèmes ne sélectionnent que des passages. Kwok et al. [Kwock, 2000], par exemple, disposent d'un moteur de recherche qui retrouve les 300 meilleurs passages de 300 à 550 mots. Cette sélection permet sans doute d'obtenir de meilleurs temps de réponses.

La recherche de la réponse, quant à elle, est réalisée à partir d'un appariement entre la question traitée et une phrase ou un extrait de phrase. Pour tous les systèmes, les éléments de base sont les mots de la question, mais on trouve ici aussi beaucoup de variantes quant aux indices supplémentaires utilisés. Kwok et al. utilisent entre autres éléments un dictionnaire de synonymes qu'ils ont constitués à partir de WordNet. Le système FALCON est différent dans la mesure où il met en œuvre un

[2] Seulement 7 jeux de réponses, sur les 42 présentés à TREC9, ont trouvé la bonne réponse au rang 1. 27 ne l'ont pas trouvée du tout.

appariement entre question et phrase fondé sur leur représentation sémantique. De plus, il est le seul système qui ajoute une justification logique des réponses trouvées.

8. Conclusion

Le but d'un système tel que QALC est de trouver la réponse à des questions factuelles sans restriction quant au domaine abordé afin que l'utilisateur n'ait plus à rechercher lui-même la réponse à sa question dans les documents sélectionnés. De plus il est important qu'un tel système donne des éléments permettant de vérifier la validité de la réponse fournie. C'est pour cette raison que nous pensons qu'il faut fournir un contexte d'une phrase au minimum. Les premiers résultats issus de l'évaluation TREC montrent que la réalisation d'un tel système est possible et que son amélioration passe par la mise en œuvre de méthodes issues du traitement automatique des langues, allant jusqu'à l'utilisation de connaissances sémantiques. L'existence d'une ressource telle que WordNet rend une telle approche envisageable à grande échelle, même si elle demande à être utilisée avec précaution.

9. Réferences

[CEL 1998] CELEX. 1998. www.ldc.upenn.edu/readme_files/celex.readme.html. Consortium for Lexical Resources, UPenn.

[FEL 1998] Fellbaum, C., editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[FER 2000] Ferret O., Grau B., Hurault-Plantet M., Illouz G., Jacquemin C. (2000), QALC — the Question-Answering system of LIMSI-CNRS, *pre-proceedings of TREC9*, NIST.

[HAR 2000] Harabagiu, S., Pasca, M., Maiorano, J.,Experiments with Open-Domain Textual Question Answering. *In Proceedings of COLING'2000* (August 2000), Saarbruken, Germany.

[ITT 2000] Ittycherian, A., Franz, M., Zhu, W-J., Ratnaparkhi, A., and Mammone, R. IBM's Statistical Question Answering System. *In TREC9 QA-Track Notebook*, NIST, Gaithersburg MD, (2000), 60-65.

[JAC 1999] Jacquemin, C. Syntagmatic and paradigmatic representations of term variation. *In Proceedings of ACL'99*, University of Maryland, 341-348.

[KWO 2000] Kwok, K.L., Grunfeld, L., Dinstl, N., and Chan, M. TREC-9 Cross Language, Web and Question-Answering Track Experiments using PIRCS. *In TREC9 QA-Track Notebook*, NIST, Gaithersburg MD, (2000), 26-35.

[LOU 1998] de Loupy C. , Bellot Patrice, El-Bèze Marc, Marteau P.-F.. Query Expansion and Classification of Retrieved Documents, *TREC* (1998), 382-389

[PRA 2000] Prager, J., Brown, E., Radev, D., and Czuba, K. One Search Engine or Two for Question-Answering. *In TREC9 QA-Track Notebook*, NIST, Gaithersburg MD, (2000), 250-254.

[SCH 1999] Schmid, H. Improvements in Part-of-Speech Tagging with an Application To German. In Armstrong S., Church K.W., Isabelle P., Manzi S., Tzoukermann E., and Yarowsky D. (eds) *Natural Language Processing Using Very Large Corpora*, Kluwer Academic Publishers, Dordrecht, 1999.