



CEA – LIST

**Laboratoire d'Ingénierie de la Connaissance
Multimédia Multilingue**

Fontenay-Aux-Roses (92)

Sujets de stages années 2008



COMMISSARIAT A L'ÉNERGIE ATOMIQUE
LIST/SRCI
Relation avec les écoles : PHILIPPE MORGANTI
Tél : 01 46 54 91 86
e-mail : philippe.morganti@cea.fr

Titres des sujets

1. Méthodes statistiques pour la reconnaissance d'entités nommées et l'adaptation rapide à un nouveau domaine	3
2. Fusion d'information texte/image pour indexer le contenu multimédia.....	5
3. Modèle d'indexation pour la manipulation de contenus multimédia (vidéos).....	7
4. Traduction automatique à l'aide d'un modèle statistique et syntaxique	9
5. Mise en relation d'entités pour l'extraction d'information	11
6. Développement d'interfaces graphiques de gestion et validation de connaissances extraites à partir de textes	14
7. Détection et reconnaissance de zones de texte dans une image.....	16
8. Indexeur de très gros volume d'images sur cluster de calcul 100 cœurs	17
9. Stratégie sélection-fusion de descripteurs images	19
10. Clusterisation automatique fondée sur des espaces vectoriels sémantiques - désambiguïsation des mots dans un corpus	20
11. Clusterisation automatique fondée sur des espaces vectoriels sémantiques - voisinage des mots	22



FICHE STAGE ≥ 4 MOIS

Sujet :

Méthodes statistiques pour la reconnaissance d'entités nommées et l'adaptation rapide à un nouveau domaine

Niveau : Bac + 5 - 3^{ème} année d'école d'ingénieur – Master

Durée du stage : 6 mois minimum

Contexte.

Le stage se situe dans le contexte de l'extraction d'information, domaine dont l'objectif est de remplir des formulaires (*templates*) prédéfinis à partir de textes. Un extrait de texte tel que :

Après trois jours de rumeurs incessantes, **Google** a annoncé lundi soir l'acquisition du site de partage de vidéos **Youtube** pour **1,65 milliard de dollars**. L'opération, approuvée par les conseils d'administration des deux sociétés, devrait être finalisée à la **fin du quatrième trimestre 2006** ...

doit ainsi permettre à un système d'extraction d'information de remplir un formulaire typique sur les rachats de sociétés qui aurait la forme suivante :

FORMULAIRE « RACHAT DE SOCIETE »

société acheteuse : Google

société achetée : Youtube

montant : 1,65 milliard de dollars

date : fin du quatrième trimestre 2006

Plus généralement, les systèmes d'extraction d'information ont pour tâche de transformer des textes en informations structurées, exploitables par exemple dans des bases de données relationnelles. Pour ce faire, ils s'appuient sur des techniques de traitement automatique des langues allant de la normalisation et du typage des mots (*a annoncé* → verbe annoncer au passé composé) jusqu'à l'explicitation de leur rôle au niveau syntaxique (*Google* est le sujet d'*annoncer*) et sémantique (*Youtube* est l'objet de l'action *acquisition*). Parmi les analyses sémantiques nécessaires, la reconnaissance des entités nommées occupe un rôle central puisqu'elle permet d'identifier les objets candidats au remplissage du formulaire considéré. Dans l'exemple ci-dessus, c'est cette reconnaissance qui permet d'identifier *Google* et *Youtube* comme des sociétés, *1,65 milliard de dollars* comme un montant financier et *fin du quatrième trimestre 2006* comme une date.

Objectifs du stage.

Les systèmes d'extraction d'information s'appuient sur des outils de traitement des langues génériques mais sont également fortement dépendants de leur domaine d'application, en particulier du fait des ressources linguistiques sur lesquelles ils s'appuient. Créer un tel système pour un nouveau domaine représente donc un investissement important en termes de main d'œuvre. Le stage s'inscrit dans une démarche visant à réduire la charge de travail inhérente à la réalisation d'un système d'extraction d'information pour un nouveau domaine. Il se focalisera plus précisément sur le problème de la reconnaissance des entités nommées caractérisant le domaine considéré.

Le laboratoire LIC2M du CEA LIST dispose déjà pour réaliser cette tâche d'un environnement de développement permettant de définir et de mettre au point manuellement des règles de reconnaissance. L'objectif du stage est de déterminer dans quelle mesure cette approche peut être complétée par des outils statistiques (classifieurs à base de maximum d'entropie, machines à vecteurs de support (SVM) ...) entraînés à partir de corpus annotés et ce, dans une perspective



d'amorçage (bootstrapping) : un premier ensemble de règles manuelles est défini pour annoter un corpus avec les entités du domaine dont la reconnaissance est la plus sûre suivant cette méthode. Ce corpus est ensuite utilisé pour entraîner un outil statistique permettant d'améliorer la couverture des entités reconnues.

Le travail comportera plus spécifiquement deux grandes phases :

- la réalisation d'un système de reconnaissance d'entités nommées de nature statistique fonctionnant par apprentissage à partir de corpus annotés ;
- l'étude de stratégies de coopération entre approche à base de règle et approche statistique dans la perspective de minimiser la charge de travail manuel pour l'adaptation à un nouveau domaine d'application.

Bibliographie

Oliver Bender, Franz Josef Och and Hermann Ney (2003) Maximum Entropy Models for Named Entity Recognition In: *Proceedings of CoNLL-2003*, Edmonton, Canada, pp. 148-151

CoNLL Shared Task: Language-Independent Named Entity Recognition (<http://www.cnts.ua.ac.be/conll2003/ner>)

Mikheev, A., Grover, C. and Moens, M. (1998). Description of the LTG system used for MUC-7. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Fairfax, Virginia.

Thierry Poibeau (2003) *Extraction d'information : du texte brut au web sémantique*. Hermès.

Domaines de spécialité requis :

Moyens informatiques mis en œuvre : Langage C++, Perl – Logiciel

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT

Labo: CEA LIST/LIC2M – Laboratoire d'Ingénierie de la Connaissance
Multimédia Multilingue

Web : <http://www->

LIST.cea.fr/fr/programmes/systemes_interactifs/systemes_interactifs.htm#1

Encadrement

Nom de l'ingénieur responsable : Olivier Ferret, Romaric Besançon **Tél.** : 01 46 54 96 45

E-mail : Olivier.ferret@cea.fr **Fax** : 01 46 54 75 80

Secrétariat : LIST/SRCI **Tél.** : 01 46 54 91 17

Nom du chef de laboratoire : Olivier Mesnard **Tél.** : 01 46 54 73 38



FICHE STAGE ≥ 4 MOIS

Sujet :

Fusion d'information texte/image pour indexer le contenu multimédia

Niveau : Bac + 5 - 3^{ème} année d'école d'ingénieur – Master 2

Durée du stage : 4 mois minimum

Résumé

Contexte: Depuis une quinzaine d'années, l'analyse automatique de documents multimédia (pages web, vidéos...) se focalise sur la description de l'information transmise par le support multimédia. La recherche et l'estimation de cette information constitue un travail d'indexation des données qui se fait actuellement essentiellement de façon manuelle.

Par nature, un document multimédia est caractérisé par une information multimodale, hétérogène et complémentaire. Il s'agit par exemple d'information textuelle, de contenu numérique provenant des images ou du canal audio (parole, musique...), ou encore d'informations de structure (spatiale, temporelle) des documents. La fusion précoce (i.e avant toute prise de décision sur chacun des canaux d'information) de ces informations est un défi difficile mais de première importance pour permettre l'indexation des documents.

Dans ce contexte, le formalisme des fonctions de croyance transférable (TBM : Transferable Belief Model) [1] offre un cadre intéressant. Il permet de fusionner des données en s'affranchissant de leur nature et de modéliser explicitement l'incertitude et le degré de confiance que l'expérimentateur place en chaque canal.

Le stage consistera à définir une architecture logicielle permettant de fusionner des informations hétérogènes (provenant principalement de « texte » et d'« image ») en s'appuyant sur le formalisme TBM.

Objectifs du stage :

Le stage s'appuiera sur les technologies existantes au laboratoire LIC2M (Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue) du CEA LIST sur l'indexation et l'analyse d'image et l'analyse linguistique et sémantique de textes. Le travail consistera à développer une architecture logicielle permettant de fusionner ces différentes briques.

Il s'agira de développer une architecture ad-hoc de fusion d'information, liée dynamiquement aux modules d'analyse des documents multimédia. Chacune de ces modules produit des informations spécifiques au canal concerné (analyse du son, de l'image, du texte...), dont le formatage sera néanmoins imposé. La brique logicielle à développer reçoit ces flux d'information sans connaître la nature profonde des informations, mais adapte le schéma de fusion en fonction d'un algorithme qui lui est propre.

Le stage comportera trois étapes classiques :

- appropriation des outils existants et état de l'art,
- définition et implémentation de l'architecture d'intégration,
- réalisation d'une interface de visualisation.

Les algorithmes utilisés seront codés en C++ (ou C).

La plate forme de référence est Linux.



Bibliographie

[1] Smets Ph. and Kennes R., *The transferable belief model*, Artificial Intelligence 66:191-234, 1994. (<http://iridia.ulb.ac.be/~psmets/AABPapers.html#A>)

Domaines de spécialité requis : Traitements de l'information, architecture logicielle, C++

Moyens informatiques mis en œuvre : Linux, C++, langages de script.

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Ingénierie de la connaissance, analyse d'image, de sons, traitement linguistique.

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT

Labo : CEA LIST/LIC2M

Web : <http://www-LIST.cea.fr/>

Encadrement

Nom de l'ingénieur responsable : Hervé Le Borgne, Romaric Besançon

Tél. : 01 46 54 85 31

E-mail : herve.le-borgne@cea.fr

Fax : 01 46 54 75 80

Secrétariat : LIST/SRCI

Tél. : 01 46 54 91 17

Nom du chef de laboratoire : Olivier MESNARD

Tél. : 01 46 54 73 38



FICHE STAGE ≥ 4 MOIS

Sujet :

Modèle d'indexation pour la manipulation de contenus multimédia (vidéos)

Niveau : Bac + 5 - 3^{ème} année d'école d'ingénieur – Master 2

Durée du stage : 4 mois minimum, 6 préférables

Résumé

Contexte: Depuis une quinzaine d'années, l'analyse automatique de documents multimédia (pages web, vidéos...) se focalise sur la description de l'information transmise par le support multimédia. La recherche et l'estimation de cette information constitue un travail d'indexation des données qui se fait actuellement essentiellement de façon manuelle.

Il existe deux types d'information attachés à un document : la structure et le contenu. L'extraction de la structure met en évidence l'organisation d'un document (e.g. un film est composé de plusieurs scènes, incluant plusieurs plans). La détermination du contenu vise à décrire les caractéristiques physiques et sémantiques des éléments de la structure (e.g. un plan contient une voiture).

Le stage consistera à définir une architecture capable d'intégrer des outils existants d'analyse de contenus multimédia en vue de les indexer de façon automatique.

Objectifs du stage

Le stage s'appuiera sur une thèse et des techniques développées au laboratoire LIC2M (Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue) du CEA LIST sur l'indexation et l'analyse de films et la classification et segmentation de données multimédia.

Le travail de stage consistera à créer une architecture logiciel en C++ afin d'intégrer les outils d'analyse développés au laboratoire. Ces outils concernent les différents canaux présents dans une vidéo: image, son, texte et transcription de parole. Le stagiaire intégrera d'une part l'extraction de certaines caractéristiques numériques et sémantiques des documents multimédias, puis la segmentation de ces documents selon une hiérarchie fixée (e.g. scène, plan, image, pour les vidéos). Plus particulièrement, on s'appuiera sur un modèle global d'indexation multimédia développé pour l'analyse de films de cinéma. L'architecture pourra s'inspirer du standard MPEG7. L'implémentation pourra bénéficier d'un cluster à haute performance de calcul en vue d'une application à l'indexation de vidéos et plus particulièrement de journaux télévisés et de documentaires.

Le stage comportera trois étapes classiques :

- appropriation des outils existants et état de l'art,
- définition et implémentation de l'architecture d'intégration,
- réalisation d'une interface de visualisation.

Les algorithmes utilisés seront codés en C++ (ou C) à partir de codes sources du laboratoire. La plate forme de référence est Linux.

Bibliographie

Modèles d'indexation multimedia pour la description automatique de films de cinéma. Bertrand Delezoide. Thèse de l'Université Pierre et Marie Curie. 2006.
<http://mediatheque.ircam.fr/articles/textes/Delezoide06c/>



Domaines de spécialité requis : Traitements de l'information, architecture logicielle, C++

Moyens informatiques mis en œuvre : Linux, C++, langages de script.

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Ingénierie de la connaissance, analyse d'image, de sons, traitement linguistique.

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT

Labo: CEA LIST/LIC2M

Web : <http://www-LIST.cea.fr/>

Encadrement

Nom de l'ingénieur responsable : Bertrand DELEZOIDE Tél. : 01 46 54 86 53

E-mail : Bertrand.delezoide@cea.fr Fax : 01 46 54 75 80

Secrétariat : LIST/SRCI Tél. : 01 46 54 91 17

Nom du chef de laboratoire : Olivier MESNARD Tél. : 01 46 54 73 38



FICHE STAGE ≥ 4 MOIS

Sujet :

Traduction automatique à l'aide d'un modèle statistique et syntaxique

Niveau : Bac + 5 - 3^{ème} année d'école d'ingénieur – Master 2

Durée du stage : 5 ou 6 mois

Résumé

Contexte: les systèmes de traduction automatique (TA) directs, interlingues ou de transfert s'appuient sur des lexiques et des règles. Dans les systèmes de TA directs, l'information est codée dans des dictionnaires. Dans les systèmes à transfert, les données sont représentées par des grammaires, des lexiques et des règles de transfert. Dans les systèmes interlingues, l'information est stockée dans des représentations sémantiques profondes.

La masse de travail nécessaire pour créer manuellement ces ressources linguistiques est importante. C'est la raison pour laquelle depuis quelques années de nombreuses tentatives d'acquisition automatique de ces informations ont été faites et des systèmes de traduction automatique basés sur des approches statistiques ont vu leur jour. Ces systèmes construisent automatiquement leurs lexiques à partir de corpus de textes bilingues (un ensemble de textes en langue source accompagnés de leur traduction en langue cible). Pour mettre en correspondance, aussi précisément que possible, chaque mot de chaque phrase d'une langue avec sa contrepartie dans l'autre langue, des algorithmes d'alignement ont été développés. Le résultat d'un tel alignement est un lexique de correspondance bilingue de mots et de phrases.

Une fois que l'algorithme d'alignement a construit le lexique, la traduction est effectuée par un procédé de substitution directe suivi d'une réorganisation des mots pour que la grammaire soit respectée. La réorganisation est réalisée à l'aide de règles dérivées de lois statistiques sur la probabilité des mots dans des contextes donnés.

Objectifs du stage

Le stage s'appuiera sur un framework open source (GenPar) permettant d'implémenter un système de traduction basé sur une approche statistique pour l'acquisition des modèles de langage et de traduction. Le framework GenPar inclut des modules d'analyse linguistique (Tokeniseur, POSTagger, Parseur, ...) et de traduction (Aligneur de mots, ...). Le Parseur de ce framework génère en sortie des arbres pondérés utilisées lors de la désambiguïsation syntaxique.

Le stage consistera d'une part, à intégrer à ce framework l'analyseur linguistique du laboratoire LIC2M (LIMA) et plus particulièrement son analyseur syntaxique (relations de dépendance), et d'autre part, à évaluer la qualité de la traduction résultat de cette intégration.

Le stage comportera les étapes suivantes :

- Appropriation des outils existants (GenPar, LIMA).
- Intégration de l'analyse syntaxique de LIMA dans GenPar.
- Evaluation du prototype de traduction français-anglais.
- Réalisation d'une interface graphique pour l'utilisation de ce prototype de traduction.

Les modules du framework GenPar sont codés en C++, Java et Perl et l'analyseur linguistique du



LIC2M est codé en C++. La plate forme de référence est Linux.

Bibliographie

- H. Schwenk, D. Déchelotte, H. Bonneau-Maynard et A. Allauzen. *Modèles statistiques enrichis par la syntaxe pour la traduction automatique*. TALN 2007, Toulouse, 5–8 juin 2007.

- A. Burbank, M. Carpuat, S. Clark, M. Dreyer, P. Fox, D. Groves, K. Hall, M. Hearne, I.D. Melamed, Y. Shen, A. Way, B. Wellington et D. Wu. *Statistical Machine Translation by Parsing*. Final report of the 2005 Language Engineering Workshop. Johns Hopkins University. November 12, 2005.

- M. Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD Dissertation, University of Pennsylvania, 1999.

Domaines de spécialité requis : Traitement Automatique de la Langue Naturelle, C++, Java, Perl.

Moyens informatiques mis en œuvre : Linux, C++, Java, Langages de script.

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Traduction automatique, Mémoire de traduction, Alignement, Traitement linguistique.

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT

Labo : CEA LIST/LIC2M

Web : <http://www-LIST.cea.fr/>

Encadrement

Nom de l'ingénieur responsable : Nasredine Semmar Tél. : 01 46 54 80 15

E-mail : nasredine.semmar@cea.fr Fax : 01 46 54 75 80

Secrétariat : LIST/SRCI Tél. : 01 46 54 91 17

Nom du chef de laboratoire : Olivier MESNARD Tél. : 01 46 54 73 38



FICHE STAGE ≥ 4 MOIS

Sujet :

Mise en relation d'entités pour l'extraction d'information

Niveau : Bac + 5 - 3^{ème} année d'école d'ingénieur – Master

Durée du stage : 6 mois minimum

Contexte.

Le stage se situe dans le contexte de l'extraction d'information, domaine dont l'objectif est d'identifier des événements ou faits dans des textes, et de structurer les informations retenues. Cela peut se traduire par le remplissage de formulaires (*templates*) prédéfinis à partir de textes. Par exemple, si l'on s'intéresse aux événements sismiques, un extrait de dépêche d'actualité tel que :

Un tremblement de terre frappe l'île indonésienne de Sumatra - *publié le lundi 12 juin 2006*

Un fort tremblement de terre a secoué **lundi** la côte occidentale de l'île indonésienne de **Sumatra**, sans qu'aucune information immédiate sur les pertes en vie humaine et en matériel ait été signalée pour l'heure.

Le séisme d'une magnitude de **5.9** était centré à **530 km de la capitale provinciale de Lampung** (sud-ouest du pays). **Banda Lampung**, a rapporté le journal Jakarta Post, citant l'Agence de Météorologie et de Géophysique. **Lampung** est située à la pointe australe de **Sumatra**, à environ **250 km au nord-ouest de Jakarta**.

Dimanche matin, deux tremblements de terre modérés ont frappé **Padang Sidempuan** et **Gunung Sitoli**, dans la province de **Sumatra Nord**.

L'épicentre du tremblement de terre d'une magnitude de **3.2** qui a frappé **Padang Sidempuan** se situait à quelque **9 km de la ville** et à environ 2,1 km de profondeur, et le séisme d'une magnitude de **4.2** qui a secoué **Gunung Sitoli**, sur **l'île de Nias**, était centré en mer à quelque **42 km au nord-ouest au large de la ville**, à 33 km de profondeur.

L'Indonésie, située dans le "Cercle d'Incendie" du Pacifique, un arc de volcans encerclant le Bassin du Pacifique, est fréquemment frappée par des tremblements de terre. Le **27 mai dernier**, un séisme d'une magnitude de **5.8** a dévasté certaines zones de la province centrale de **Java** et de la région de **Yogyakarta**, **tuant plus de 5 800 personnes**.

doit ainsi permettre à un système d'extraction d'information d'identifier les 4 événements sismiques différents suivants :

lieu
date
magnitude
dégâts

Sumatra
12/06/06
5.9

Padang Sidempuan
11/06/06
3.2

Gunung Sitoli
11/06/06
4.2

Java, Yogyakarta
25/05/06
5.9



tuant plus de 5800 personnes

Les systèmes d'extraction d'information s'appuient sur des techniques de traitement des langues, permettant l'analyse des phrases, au niveau syntaxique et sémantique. Ces traitements permettent en particulier l'identification d'entités spécifiques au domaine concerné (dans l'exemple, lieu, date, magnitude etc), et de leurs relations, puis la mise en correspondance de ces éléments dans des cadres plus larges (*templates* ou *scenarios*).

Objectifs du stage.

Les systèmes d'extraction d'information s'adaptent à des domaines d'application différents en les paramétrant avec de ressources linguistiques spécifiques au domaine. C'est le cas par exemple de l'extraction des entités spécifiques pour les événements sismiques. Le développement de ces ressources peut être manuel ou automatique.

En ce qui concerne les stratégies d'agrégation des éléments en *templates*, plusieurs approches ont été testées, pouvant utiliser des algorithmes d'apprentissage statistique ou des grammaires locales. Néanmoins, pour traiter des cas complexes comme l'exemple précédent, dans lequel plusieurs événements à repérer sont imbriqués dans le même texte, une analyse de la structure du discours peut être nécessaire.

L'objectif du stage est de développer les techniques de regroupement des entités spécifiques repérées pour le remplissage de *templates*. L'idée est de s'appuyer d'abord sur des relations locales entre les entités, en fonction de leur contexte, puis d'utiliser les relations repérées pour remplir les *templates*, en s'appuyant sur l'utilisation jointe d'heuristiques de bas niveau (proximité des entités), avec l'exploitation de l'analyse linguistique profonde, en particulier les relations syntaxiques et la structure du discours. Les techniques développées devront être suffisamment génériques pour pouvoir s'appliquer à différents domaines.

Le laboratoire LIC2M du CEA LIST dispose d'une plate-forme modulaire de traitement des langues qui permet de réaliser l'analyse linguistique d'un texte, jusqu'au niveau syntaxique au moins. Cette plate-forme dispose en particulier d'un moteur générique d'application de patrons morpho-syntaxiques permettant l'extraction d'entités ou de relations à partir de règles (écrites manuellement). Le stagiaire travaillera à partir d'un prototype d'extraction d'information développé sur le domaine des événements sismiques et l'étendra pour rendre plus génériques les stratégies de regroupement des entités pour remplir les *templates*, dans le but de pouvoir les appliquer à d'autres domaines (par exemple, pour la veille économique et financière : repérer les événements financiers importants – fusion/acquisition).

Bibliographie

Thierry Poibeau (2003) *Extraction d'information : du texte brut au web sémantique*. Hermès.

Douglas E. Appelt, David J. Israel (1999) *Introduction to Information Extraction Technology* Tutorial Prepared for IJCAI-99

K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, Y. Wilks (1998) University of Sheffield: Description of the LaSIE-II System used for MUC-7. *Proceedings of the MUC-7 Message Understanding Conference*

S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, R. Weischedel (1998) Algorithms that learn to extract information. BBN: Description of the SIFT system as used for MUC-7. *Proceedings of the*



COMMISSARIAT A L'ENERGIE ATOMIQUE
LIST/SRCI
Relation avec les écoles : PHILIPPE MORGANTI
Tél : 01 46 54 91 86
e-mail : philippe.morganti@cea.fr

MUC-7 Message Understanding Conference

Domaines de spécialité requis :

Moyens informatiques mis en œuvre : Langage C++, Perl – Logiciel

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Exemple : Plateforme
Participation à un projet européen, industriel,

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT

Labo : CEA LIST/LIC2M – Laboratoire d'Ingénierie de la Connaissance
Multimédia Multilingue
Web : <http://www-LIST.cea.fr/>

Encadrement

Nom de l'ingénieur responsable :	Romarc BESANCON	Tél. :	01 46 54 80 16
E-mail :	Romarc.besancon@cea.fr	Fax :	01 46 54 75 80
Secrétariat :	LIST/SRCI	Tél. :	01 46 54 91 17
Nom du chef de laboratoire :	Olivier Mesnard	Tél. :	01 46 54 73 38



COMMISSARIAT A L'ENERGIE ATOMIQUE
LIST/SRCI
Relation avec les écoles : PHILIPPE MORGANTI
Tél : 01 46 54 91 86
e-mail : philippe.morganti@cea.fr

FICHE STAGE \geq 4 MOIS

Sujet :

Développement d'interfaces graphiques de gestion et validation de connaissances extraites à partir de textes

Niveau : Bac + 5 - 3^{ème} année d'école d'ingénieur – Mastère

Durée du stage : 4 à 6 mois

Résumé

Le laboratoire LIC2M développe des outils d'enrichissement et de peuplement d'ontologies. Une ontologie est une formalisation des connaissances d'un domaine particulier. Ce type de modèle est indispensable pour effectuer l'annotation sémantique de textes et une extraction d'information ciblée pour de la veille par exemple. La construction manuelle d'une ontologie est un processus très coûteux qui nécessite une grande expertise du domaine traité ainsi que des ontologies elles-mêmes. Ces connaissances sont présentes au moins partiellement dans les textes du domaine et peuvent être extraites à l'aide d'outils de traitement automatique des langues. Les connaissances extraites, souvent bruitées, nécessitent une validation par l'expert via des interfaces graphiques adaptées. L'objectif du stage est de développer ces interfaces pour accélérer la phase de paramétrage et de constitution de ressources pour des outils de veille.

Objectifs du stage

Le travail du stagiaire consistera à

- Etudier les interfaces graphiques de gestion d'ontologies existantes dans le domaine du logiciel libre, en particulier Protégé et ses plug-ins de validation de connaissances et Terminae.
- Concevoir les interfaces et les méthodes d'interaction avec l'expert sachant qu'elles devront systématiquement permettre le retour au texte (affichage des occurrences textuelles correspondant à une ou plusieurs connaissances extraites).
- Développer, tester et valider les interfaces.

Le stage comportera trois parties :

- état de l'art,
- développement et documentation,
- test et validation des développements réalisés.

Selon les résultats, possibilité de proposition de publication.

Domaines de spécialité requis : des connaissances en Interfaces Graphiques sont nécessaires et si possible en Qt/KDE. La connaissance de Java sera utile.

Moyens informatiques mis en œuvre : Langage C++, Bibliothèques Qt et KDE

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Exemple : Outils d'extraction de connaissances développés au sein du laboratoire

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT

Labo : CEA LIST/LIC2M – Laboratoire d'Ingénierie de la Connaissance
Multimédia Multilingue
Web : <http://www-LIST.cea.fr/>



COMMISSARIAT A L'ENERGIE ATOMIQUE
LIST/SRCI
Relation avec les écoles : PHILIPPE MORGANTI
Tél : 01 46 54 91 86
e-mail : philippe.morganti@cea.fr

Encadrement

Nom de l'ingénieur responsable : Faiza Gara	Tél. : 01 46 54 80 58
E-mail : Faiza.Gara@cea.fr	Fax : 01 46 54 75 80
Secrétariat : LIST/SRCI	Tél. : 01 46 54 91 17
Nom du chef de laboratoire : Olivier Mesnard	Tél. : 01 46 54 73 38



COMMISSARIAT A L'ENERGIE ATOMIQUE
LIST/SRCI
Relation avec les écoles : PHILIPPE MORGANTI
Tél : 01 46 54 91 86
e-mail : philippe.morganti@cea.fr

FICHE STAGE ≥ 4 MOIS

Sujet :

Détection et reconnaissance de zones de texte dans une image

Niveau : Bac + 5 - 3^{ème} année d'école d'ingénieur – Master

Durée du stage : 4 mois minimum, 6 préférable

Résumé.

Contexte :

Le LIC2M a été créé en février 2002 au sein du CEA LIST (DRT/LIST/SRCI), au CEA Fontenay-aux-Roses. Il rassemble des compétences en ingénierie linguistique et en traitement d'images. L'activité du laboratoire porte sur l'exploitation de l'information exprimée dans différents médias (principalement image fixes ou vidéo et texte) pour produire de la connaissance. Le laboratoire est ainsi spécialisé dans l'indexation d'images et le traitement d'informations textuelles multilingues, il est ainsi structuré en deux « pôles » travaillant en étroite coopération. Le sujet de ce stage est proposé par le pôle image.

Objectifs du stage.

Les indications textuelles présentes dans une image sont des informations importantes pour l'extraction automatique de connaissances. Ces informations permettent en effet d'apporter des connaissances sur le contenu et/ou le contexte d'une image (nom de marque, panneau de signalisation...); connaissances permettant de faciliter l'indexation automatique des images. L'extraction de ces connaissances implique deux phases. Une phase de détection visant à localiser et dénombrer les zones de texte dans l'image et une phase de reconnaissance visant à extraire – et normaliser – le texte (OCR). Le stage comportera trois étapes: état de l'art, recherche, test et validation des méthodes employées. Les algorithmes utilisés seront codés en C++ (ou C) à partir de codes sources du laboratoire.

Domaines de spécialité requis :

Moyens informatiques mis en œuvre : Langages : C et C++, OS Windows et Linux

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Traitement d'images : segmentation, filtrage, morphologie mathématique, texture, forme

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT

Labo : CEA LIST/LIC2M – Laboratoire d'Ingénierie de la Connaissance
Multimedia Multilingue

Web : <http://www-LIST.cea.fr/>

Encadrement

Nom de l'ingénieur responsable : Pierre-Alain MOELLIC	Tél. : 01 46 54 96 19
E-mail : Pierre-alain.moellic@cea.fr	Fax : 01 46 54 75 80
Secrétariat : LIST/SRCI	Tél. : 01 46 54 91 17
Nom du chef de laboratoire : Olivier MESNARD	Tél. : 01 46 54 73 38



FICHE STAGE ≥ 4 MOIS

Sujet :

Indexeur de très gros volume d'images sur cluster de calcul 100 cœurs

Niveau : Bac + 5 - 3^{ème} année d'école d'ingénieur – Master

Durée du stage : 4 mois minimum

Résumé.

Le LIC2M a été créé en février 2002 au sein du CEA LIST (DRT/LIST/SRCI), au CEA Fontenay-aux-Roses. Il rassemble des compétences en ingénierie linguistique et en traitement d'images. L'activité du laboratoire porte sur l'exploitation de l'information exprimée dans différents médias (principalement image fixes ou vidéo et texte) pour produire de la connaissance. Le laboratoire est ainsi spécialisé dans l'indexation d'images et le traitement d'informations textuelles multilingues, il est ainsi structuré en deux « pôles » travaillant en étroite coopération. Le sujet de ce stage est proposé par le pôle image.

Objectifs du stage.

L'objectif est de mettre en place un ensemble d'algorithmes paramétrés et de déterminer par calcul énumératif les solutions les meilleurs sur plusieurs bases d'images. Le but étant d'implémenter un descripteur à signature très courte pour indexer de gros volume. En effet, l'indexation couleur par exemple peut être réalisée avec plus d'une dizaine de modèles associés à des métriques ou distances en nombre au moins égal. A l'aide des moyens de calcul du laboratoire (cluster de 100 cœurs de calcul) on implémentera des méthodes génériques pouvant être interchangeables de façon à balayer l'ensemble des paramètres possibles pour un ou plusieurs indexeurs couleur, texture et/ou forme.

L'objectif de ce stage est d'intégrer ces fonctions dans le moteur image déjà existant pour créer un nouvel indexeur performant. Le stage s'appuiera sur les algorithmes d'indexation du laboratoire et, pour le calcul de la précision rappel associée, il utilisera la vérité terrain existante pour chaque base d'images.

Le stage comportera trois étapes classiques : état de l'art, recherche, test et validation des méthodes employées. Les algorithmes utilisés seront codés en C++ (ou C) à partir de codes sources du laboratoire.

Domaines de spécialité requis :

Moyens informatiques mis en œuvre : Langages : C et C++, Windows et Linux

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Connaissance théorique : traitement et indexation d'images, segmentation, filtrage, morphologie mathématique, texture, forme, notion de techniques d'apprentissage.

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT

Labo : CEA LIST/LIC2M – Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue

Web : <http://www-LIST.cea.fr/>



Encadrement

Nom de l'ingénieur responsable : Patrick Hède Tél. : 01 46 54 96 46

E-mail : patrick.hede@cea.fr Fax : 01 46 54 75 80

Secrétariat : LIST/SRCI Annick LATARE Tél. : 01 46 54 91 17

Nom du chef de laboratoire : Olivier MESNARD Tél. : 01 46 54 73 38



Sujet :

Stratégie sélection-fusion de descripteurs images

Niveau : Bac + 5 - 3^{ème} année d'école d'ingénieur – Master

Durée du stage : 4 mois minimum

Résumé.

Contexte : Le LIC2M a été créé en février 2002 au sein du CEA LIST (DRT/LIST/SRCI), au CEA Fontenay-aux-Roses. Il rassemble des compétences en ingénierie linguistique et en traitement d'images. L'activité du laboratoire porte sur l'exploitation de l'information exprimée dans différents médias (principalement image fixes ou vidéo et texte) pour produire de la connaissance. Le laboratoire est ainsi spécialisé dans l'indexation d'images et le traitement d'informations textuelles multilingues, il est ainsi structuré en deux « pôles » travaillant en étroite coopération. Le sujet de ce stage est proposé par le pôle image.

Objectifs du stage.

Le laboratoire dispose de son propre moteur de recherche d'images par le contenu PiRiA (http://www-LIST.cea.fr/fr/programmes/systemes_interactifs/labo_lic2m/piria/w3/pirianet.php). Plus d'une dizaine de descripteurs globaux et locaux existent. Le but est de mettre en place des stratégies de détermination automatique du ou des meilleurs descripteurs à utiliser pour une image requête donnée (pour une base d'images fixée) sans exclure les méthodes plus classiques de pondération afin de maximiser la qualité des réponses. Le laboratoire dispose de plusieurs bases avec leur vérité terrain de manière à juger qualitativement de la pertinence des méthodes utilisées. Lors du stage on veillera à la possibilité de généraliser la méthode en association avec le média textuel. Les développements seront à réaliser de manière préférentielle sous Linux en C++.

Domaines de spécialité requis :

Moyens informatiques mis en œuvre : Langages : C et C++, OS Windows et Linux

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Connaissance théorique en analyses, traitements et indexation d'images, segmentation, filtrage, méthodes statistiques, texture, forme. Recommandé : notions de techniques d'apprentissage.

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT

Labo : CEA LIST/LIC2M – Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue

Web : <http://www-LIST.cea.fr/>

Encadrement

Nom de l'ingénieur responsable : Parick Hède Tél. : 01 46 54 96 46

E-mail : Patrick.hede@cea.fr Fax : 01 46 54 75 80

Secrétariat : LIST/SRCI annick LATARE Tél. : 01 46 54 91 17

Nom du chef de laboratoire : Olivier MESNARD Tél. : 01 46 54 73 38



Sujet :

Clusterisation automatique fondée sur des espaces vectoriels sémantiques - désambiguïsation des mots dans un corpus

Niveau : Bac + 5 - 3^{ème} année d'école d'ingénieur – Master

Durée du stage : 4 mois minimum

Résumé.

Les espaces vectoriels sémantiques comme ceux produits par la "Latent Semantic Analysis" permettent d'obtenir une projection d'une langue sous une forme numérique. Cette projection permet d'en extraire des informations de proximité ou similarité qui évoquent fortement un lien avec une certaine sémantique, même si cette sémantique n'est pas encore très bien définie. Plusieurs voies ont été explorées pour améliorer la qualité de l'information sémantique ainsi extraite :

- soit en améliorant la méthode utilisée pour compresser et lisser la matrice initiale (on passe classiquement d'une cinquantaine de milliers de dimensions à quelques centaines),
- soit en faisant varier l'information initiale utilisée pour construire la matrice initiale. La technique traditionnelle est de parcourir le corpus et de compter les mots qui apparaissent dans une fenêtre de N mots autour du mot courant. D'autres méthodes prennent en compte des informations syntaxiques ou bien la présence de certains mots outils entre deux mots (Latent Relational Analysis).

Objectifs du stage.

La méthode que nous souhaitons explorer consiste à améliorer la qualité du corpus à partir duquel est construit l'espace vectoriel sémantique, en réduisant la polysémie des termes. En effet, lorsqu'un terme donné est polysémique (par exemple le nom commun "sens" peut signifier 1/ direction 2/ signification 3/ capacité sensorielle 4/ faculté intuitive 5/ faculté de raisonnement 6/ point de vue), sa position dans l'espace sémantique sera nécessairement un compromis entre ses différents sens, ce qui ne peut qu'impliquer une moindre qualité de cet espace sémantique (dans ce cas, la polysémie introduit un bruit dans l'espace sémantique). Or il est justement possible d'utiliser les espaces vectoriels sémantiques pour désambiguïser les mots (en étudiant, dans cet espace, les groupements de mots à proximité du mot à désambiguïser).

Ce stage concerne la seconde étape de ce sujet de recherche, à savoir la désambiguïsation automatique des mots dans un corpus. Il faudra dans un premier temps s'approprier les corpus et les outils servant à en extraire un espace vectoriel, désambiguïser manuellement une petite sous-partie du corpus, puis concevoir, développer et évaluer un outil permettant d'affecter automatiquement un de ses sens à un mot en fonction du contexte dans lequel il apparaît dans le corpus. A terme, l'outil devra se baser sur le résultat d'une clusterisation automatique et être appliqué sur un corpus entier.

Bibliographie

S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.



Y. Zhao and G. Karypis. 2002. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the 11th Conference of Information and Knowledge Management (CIKM)*, pages 515–524.

H. Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.
T. Pedersen. 2006. Unsupervised Corpus Based Methods for WSD (Pedersen), In Agirre, E. and Edmonds, P. (Editors), *Word Sense Disambiguation : Algorithms and Applications*, June 2006, pp. 133-166, Springer.

Domaines de spécialité requis : Traitement de l'information, Architecture logicielle

Moyens informatiques mis en œuvre : C/C++, environnement GNU, Cluster 96 cœurs

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Outils d'analyse de corpus, analyse numérique, Semantic Map (ressource développée au LIC2M)

Centre : Fontenay-aux-Roses **Pôle ou Direction :** DRT

Labo : CEA LIST/LIC2M
Web : <http://www-LIST.cea.fr/>

Encadrement

Nom de l'ingénieur responsable :	Gregory Grefenstette	Tél. :	01 46 54 96 56
E-mail :	gregory.grefenstette@cea.fr	Fax :	01 46 54 75 80
Secrétariat :	LIST/SRCI	Tél. :	01 46 54 91 17
Nom du chef de laboratoire :	Olivier Mesnard	Tél. :	01 46 54 73 38



FICHE STAGE ≥ 4 MOIS

Sujet :

Clusterisation automatique fondée sur des espaces vectoriels sémantiques - voisinage des mots

Niveau : Bac + 5 - 3^{ème} année d'école d'ingénieur – Master

Durée du stage : 4 mois minimum

Résumé.

Les espaces vectoriels sémantiques comme ceux produits par la "Latent Semantic Analysis" permettent d'obtenir une projection d'une langue sous une forme numérique. Cette projection permet d'en extraire des informations de proximité ou similarité qui évoquent fortement un lien avec une certaine sémantique, même si cette sémantique n'est pas encore très bien définie. Plusieurs voies ont été explorées pour améliorer la qualité de l'information sémantique ainsi extraite :

- soit en améliorant la méthode utilisée pour compresser et lisser la matrice initiale (on passe classiquement d'une cinquantaine de milliers de dimensions à quelques centaines),
- soit en faisant varier l'information initiale utilisée pour construire la matrice initiale. La technique traditionnelle est de parcourir le corpus et de compter les mots qui apparaissent dans une fenêtre de N mots autour du mot courant. D'autres méthodes prennent en compte des informations syntaxiques ou bien la présence de certains mots outils entre deux mots (Latent Relational Analysis).

Objectifs du stage.

La méthode que nous souhaitons explorer consiste à améliorer la qualité du corpus à partir duquel est construit l'espace vectoriel sémantique, en réduisant la polysémie des termes. En effet, lorsqu'un terme donné est polysémique (par exemple le nom commun "sens" peut signifier 1/ direction 2/ signification 3/ capacité sensorielle 4/ faculté intuitive 5/ faculté de raisonnement 6/ point de vue), sa position dans l'espace sémantique sera nécessairement un compromis entre ses différents sens, ce qui ne peut qu'impliquer une moindre qualité de cet espace sémantique (dans ce cas, la polysémie introduit un bruit dans l'espace sémantique). Or il est justement possible d'utiliser les espaces vectoriels sémantiques pour désambigüiser les mots (en étudiant, dans cet espace, les groupements de mots à proximité du mot à désambigüiser).

Ce stage concerne la première étape de ce sujet, à savoir la clusterisation automatique du voisinage des mots. Il faudra dans un premier temps s'appropriier les corpus et les outils servant à en tirer un espace vectoriel, puis concevoir, développer et évaluer un outil permettant de faire ressortir les clusters sémantiques au voisinage d'un mot donné. A terme, l'outil devra être appliqué à l'ensemble du lexique et servira pour effectuer une désambigüisation du corpus initial.

Bibliographie

S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.



COMMISSARIAT A L'ENERGIE ATOMIQUE
LIST/SRCI
Relation avec les écoles : PHILIPPE MORGANTI
Tél : 01 46 54 91 86
e-mail : philippe.morganti@cea.fr

Y. Zhao and G. Karypis. 2002. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the 11th Conference of Information and Knowledge Management (CIKM)*, pages 515–524.

H. Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

T. Pedersen. 2006. Unsupervised Corpus Based Methods for WSD (Pedersen), In Agirre, E. and Edmonds, P. (Editors), *Word Sense Disambiguation : Algorithms and Applications*, June 2006, pp. 133-166, Springer.

Domaines de spécialité requis : Traitement de l'information, Architecture logicielle

Moyens informatiques mis en œuvre : C/C++, environnement GNU, Cluster 96 cœurs

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Outils d'analyse de corpus, analyse numérique, Semantic Map (ressource développée au LIC2M)

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT

Labo : CEA LIST/LSI – Laboratoire de Simulation Interactive

Web : <http://www-LIST.cea.fr/>

Encadrement

Nom de l'ingénieur responsable : Gregory Grefenstette Tél. : 01 46 54 96 56

E-mail : gregory.grefenstette@cea.fr Fax : 01 46 54 75 80

Secrétariat : LIST/SRCI Tél. : 01 46 54 91 17

Nom du chef de laboratoire : Olivier Mesnard Tél. : 01 46 54 73 38



COMMISSARIAT A L'ENERGIE ATOMIQUE
LIST/SRCI
Relation avec les écoles : PHILIPPE MORGANTI
Tél : 01 46 54 91 86
e-mail : philippe.morganti@cea.fr

FICHE STAGE ≥ 4 MOIS