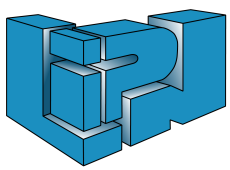
Annotation linguistique de documents Web dans une architecture distribuée et adaptable

Julien Derivière, Thierry Hamon

LIPN – UMR CNRS 7030
99 av. J.B. Clément, F-93430 Villetaneuse, FRANCE
Tél. : 33 1 49 40 28 32, Fax. : 33 1 48 26 07 12
prenom.nom@lipn.univ-paris13.fr
http://www-lipn.univ-paris13.fr/~nom

Journées Perl 2006



Introduction

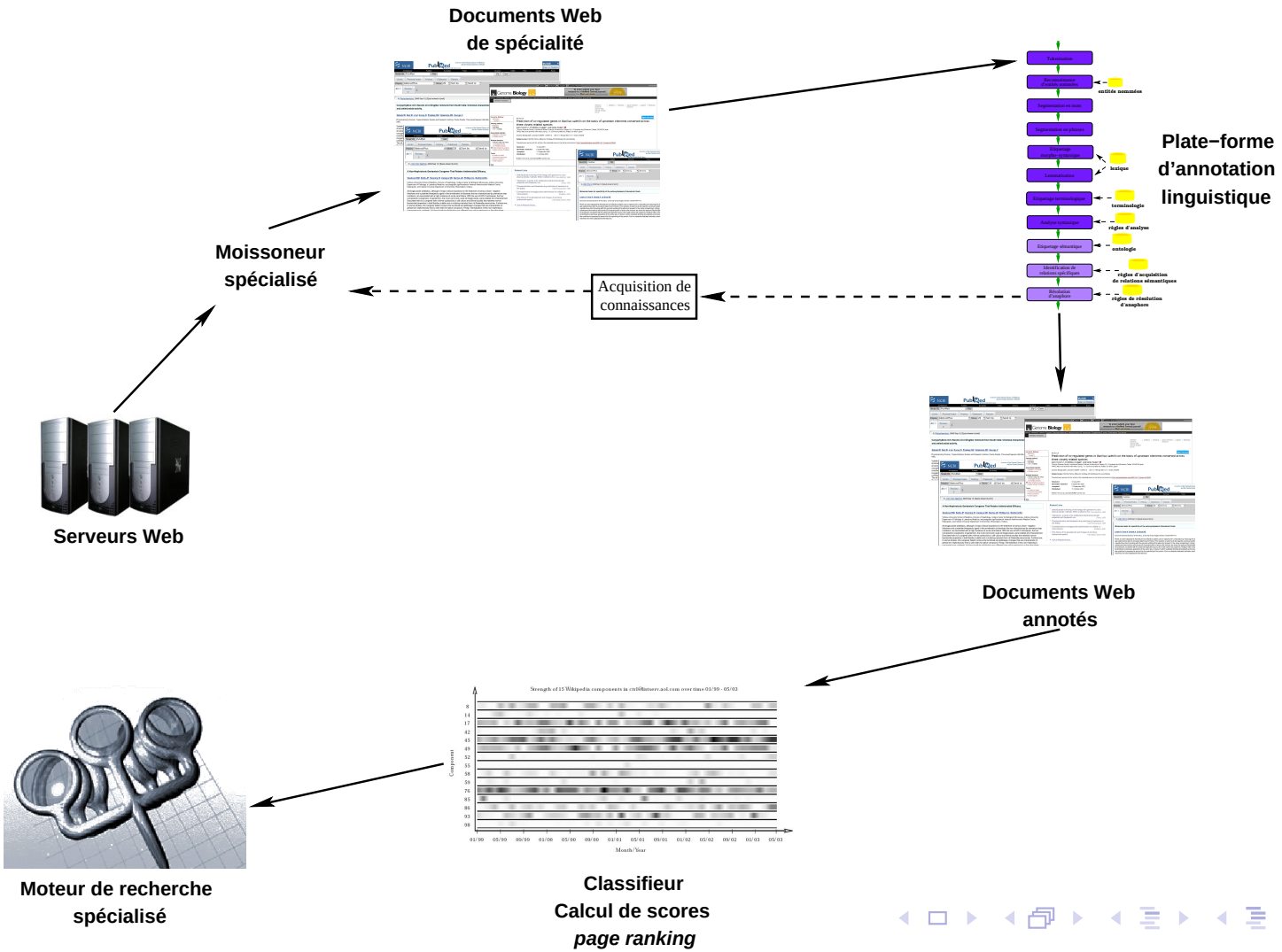
Projet ALVIS (Projet Européen STREP IST-1-002068-STP, `http://www.alvis.info/alvis/`, 2004-2006):

- développement de moteurs de recherche spécialisés, basés sur la technologie peer-to-peer
- **intégration d'informations linguistiques pour prendre en compte la spécificité du contenu des documents spécialisés**

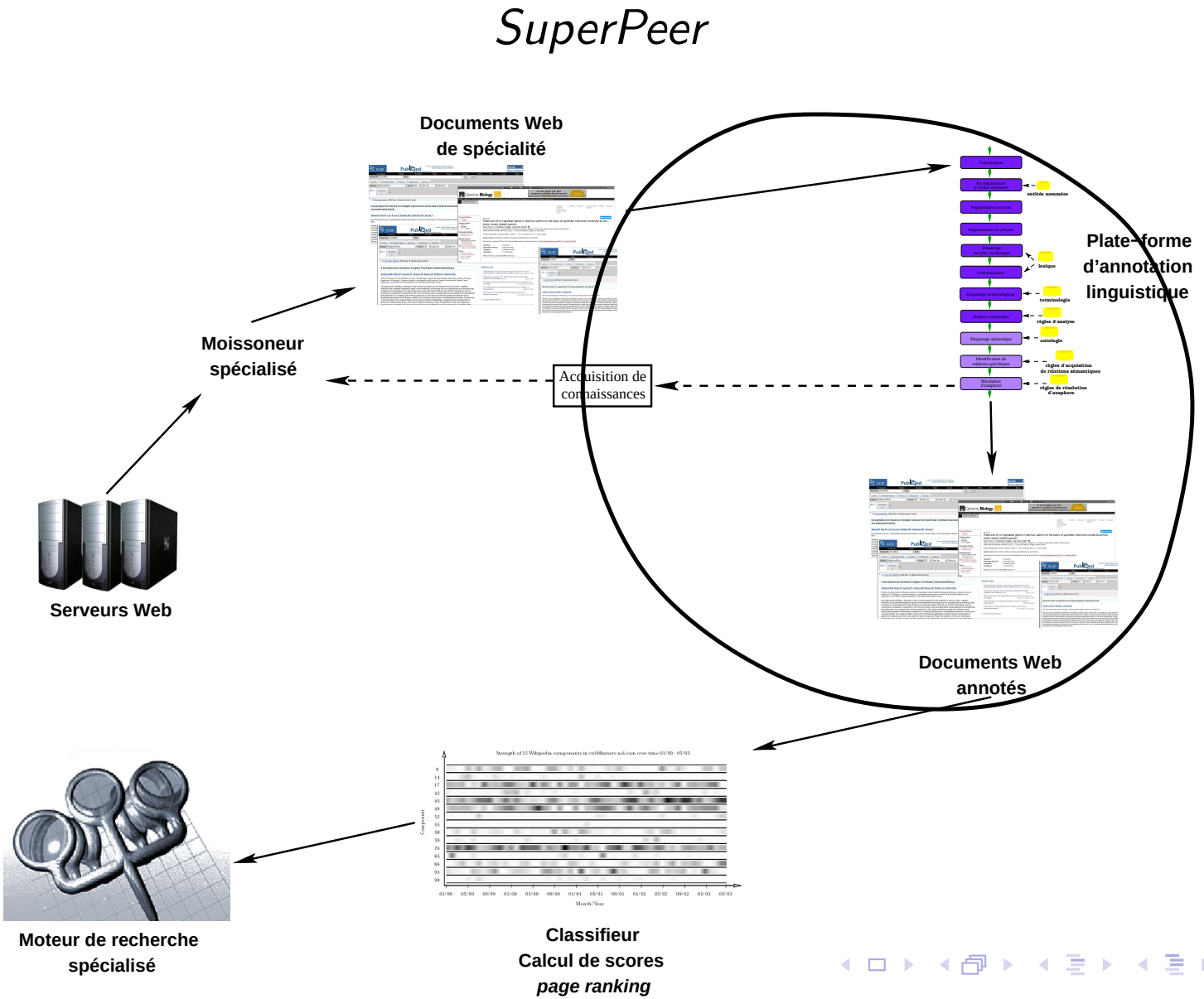
$\Rightarrow$ *concevoir et développer une architecture de Traitement Automatique de la Langue (TAL) pour annoter des documents issus du web avec des informations linguistiques*

Place du TAL dans la pipeline Alvis

SuperPeer



Place du TAL dans la pipeline Alvis



Annotation linguistique de documents

identification et description des segments de textes

→ *Enrichir les documents en associant des informations linguistiques aux différents segments de texte*

- Unités textuelles : mots, phrases, syntagmes (groupes nominaux et verbaux), unités sémantiques (entités nommées, termes)
- Propriétés morpho-syntaxiques :
  - lemmes (forme canonique) : **proteins**/*protein*
  - traits morpho-syntaxiques (genre, nombre, catégorie grammaticale) : **proteins**/*nom pluriel*
  - relations syntaxiques (sujet, objet, modifieur, etc.)
- Propriétés sémantiques :
  - catégories sémantiques : **SpoIIID**/*facteur de transcription*, **cotX**/*gene*, **Bacillus subtilis**/*espèce*
  - relations sémantiques (spécifiques au domaine, anaphore) : **GerE** `est en interaction avec` **cotB**

Quels enjeux pour le TAL ?

- Annotation de gros volumes de documents issus du Web (plusieurs dizaines de millions de mots), en un temps raisonnable
  (habituellement, sur des domaines spécialisés, quelques centaines de milliers de mots)

  $\longrightarrow$ *Comment agencer l'analyse linguistique pour obtenir un processus robuste et adapté au volume de données à traiter ?*

- Intégration d'informations linguistiques dans un moteur de recherche : aucune conclusion précise sur leur utilité

  $\longrightarrow$ Quel est l'apport d'informations linguistiques dans un domaine spécialisé ?

Plate-forme de TAL pour l'annotation de documents web

Spécifications :

- Simple d'utilisation
- Générique afin de pouvoir traiter des documents spécialisés, et l'adapter à de nouveaux domaines
- Analyse du texte rapide et robuste (*Contrainte non-traditionnelle en TAL*)
- Capable d'annoter de grandes collections de documents (plusieurs dizaines de millions de mots)
- Gestion de l'hétérogénéité des documents (langue, taille et contenu)
- Facilité d'adaptation et d'intégration de nouveaux outils

Plate-forme de TAL Alvis

- Exploitation d'outils de TAL existants

→ Ajustement/adaptation de outils au domaine traité à l'aide de ressources supplémentaires spécifiques au domaine

→ Définition d'une architecture modulaire et distribuée

  - Encapsulation des outils

  → Utilisation du modèle client/serveur

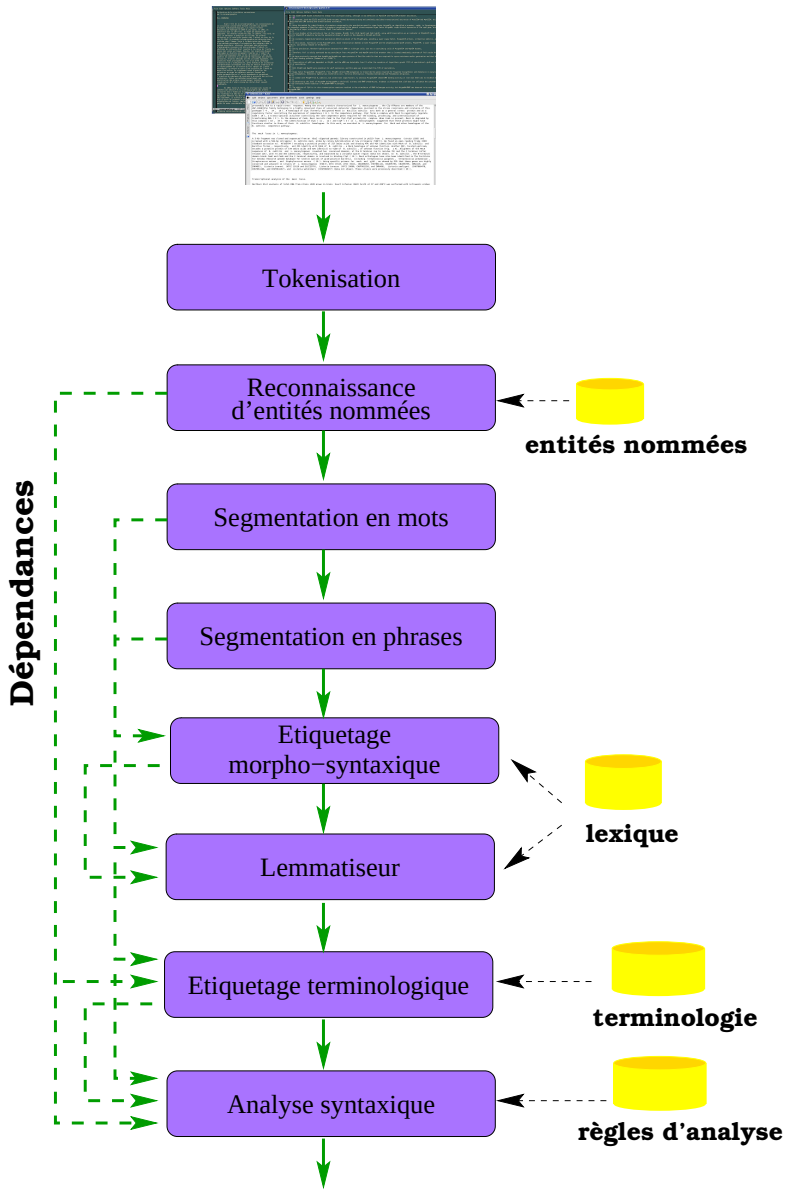
Contraintes spécifiques

- Ingénierie logicielle
  - Hétérogénéité des formats d'entrées/sortie des outils intégrés
  - Définition d'un format d'échange suffisamment générique pour ne pas refléter les sorties d'un outil en particulier
- Analyse linguistique de documents spécialisés
  - Prise en compte des informations calculées précédemment
  - Disponibilité des ressources lexicales et ontologiques
  - Temps de calcul de certains traitements (analyse syntaxique)

Une plate-forme modulaire

- Division de l'analyse linguistique en modules
- Annotations stockée dans un format XML déporté (codage UTF-8, DTD Alvis)
- Encapsulation des outils de TAL afin d'assurer de la conformité des formats d'entrées/sorties
  $\Longrightarrow$ Substitution d'outils sans impact sur l'architecture
- Adaptation au domaine assurée par les ressources utilisées par chaque module
  une liste de nom de gènes ou d'espèces est ajoutée au module de reconnaissance d'entités nommées pour traiter des textes de biologie

Description de la plate-forme et des modules



- En entrée : documents web (html, xml, Doc, PDF, postscript, LaTeX, RTF, texte, etc.) déjà téléchargés, nettoyés, encodés en UTF-8, formatés en XML
- Tokenization : Définition d'adresses pour les unités linguistiques, typage des chaînes de caractères (`B/./ /subtilis/`)
- Reconnaissance d'entités nommées et typage (*Bacillus subtilis*, *SpoIIID*)
  Place de la reconnaissance d'entités nommées : très tôt dans la plate-forme
- Etiquetage de termes (*gene expression*, *spore coat cell*)
- Analyse syntaxique : relations de dépendance entre les mots d'une phrase
  Point critique: temps de traitement important suivant les outils

Implémentation
Caractéristiques générales

Deux types d'utilisation :

- autonome
- client/server
  - Client : Traitements linguistiques des documents envoyés par le serveur
  - Serveur : Réception des documents venant du crawler, envoi des documents aux clients, envoi des documents annotés à la suite de la pipeline Alvis

  Le protocole assure qu'aucun document n'est perdu

Langues : Anglais et français
Implementation en Perl, sous GPL
Disponible comme modules CPAN

Implémentation
Description

- Trois modules principaux (`Alvis::NLPPlatform`, `Alvis::NLPPlatform::UserNLPWrappers`, `Alvis::NLPPlatform::NLPWrappers`)
- Scripts pour les modes client/serveur et autonome (`alvis-nlp-server`, `alvis-nlp-client`, `alvis-nlp-standalone`)
- Fichiers de configuration :
  - Paramètres de la plate-forme
  - Chemin et ligne de commande facilitant l'exécution des outils
  - Définition des ressources

Implémentation
Utilisation et adaptation

- Utilisation simple : exploitation de *wrappers* par défaut
  Ajustement des paramètres à l'environnement local

- Adaptation au domaine :
  Définition des ressources pour l'adaptation au domaine spécialisé dans le fichier de configuration

- Intégration de nouveaux outils :
  Masquage du module `Alvis::NLPPlatform::UserNLPWrappers`
  - Définition des nouveaux *wrappers* dans un module `Alvis::NLPPlatform::UserNLPWrappers` local à l'utilisateur
  - Définition de l'accès prioritaire au module `Alvis::NLPPlatform::UserNLPWrappers` local par modification de la variable `PERL5LIB`

Analyse des performances

Matérial

- Experience sur une collection de 55,329 documents (plus de 80 millions de mots)
  domaine : biologie
- Taille des documents XML : entre 1Ko et 100Ko
- Taille du plus gros document : 5.7 Mo
- Annotation jusqu'à l'étiquetage de termes

$\longrightarrow$ Indication des temps de traitement de la plate-forme et de son utilisabilité sur de gros volumes de documents

Analyse des performances

Machines

Annotation avec 16 machines :

- Principalement : PC standard avec 1Go de RAM et un processeur de 3.1 ou 2.9 GHz
  un client par machine avec une priorité faible
- 4 éléments d'un cluster avec une configuration similaire
  un client par élément avec une priorité faible
- Machine avec 8GB de RAM et deux processeurs Xeon 2.8GHz Xeon (dual-core)
  serveur et trois clients

Système d'exploitation : Linux Debian ou Mandrake

Résultats

Annotation distributée : trois jours (traitement séquenciel: 25 jours et 20h39'23")

En moyenne 2790 documents par client

Temps de traitement moyen par document : 1 minute

Taille du plus gros document traité : 350,444 mots

27 documents non annotés (0.04%)

Difficultés

- Suspension de clients : bug dans l'un des outils de TAL (augmentation des temps de traitement)

  $\longrightarrow$ Correction apportée après l'expérience

  Approximation du temps de traitement global : 2 jours et 7 heures (temps de traitement du client le plus lent)
- PC standard en utilisation normale : variation dans la charge du CPU et redémarrage de machines
- Au cours du développement : difficultés avec l'utilisation du jeu de caractères UTF-8 avec certains outils de TAL, et sur différentes machines et environnements

Conclusion

Conception d'une plate-forme d'annotation linguistique :

- Utilisation d'outils existants
- Annotation de gros volume de textes : distribution des traitements linguistiques sur plusieurs machines
- Robustesse au niveau des clients et des composants
- Flexibilité de la plate-forme pour la configuration et l'intégration d'outils supplémentaires

Perspectives

- Intégration d'autres étapes (étiquetage sémantique, identification de relations sémantiques et d'anaphore)
- Définition de modes d'exécution : acquisition et production
- Gestion de la distribution : traitement de documents avec équilibrage de charge (envoi d'un document à un client en fonction de ses capacités à effectuer par le traitement)
- Améliorations à apporter au niveau des temps de traitements et du code


Berroyer (Jean-François). –
TagEN, un analyseur d"entités nommées : conception, développement et évaluation. –
Mémoire de d.e.a. d'intelligence artificielle, Université Paris-Nord, 2004.


Grefenstette (G.) et Tapanainen (P.). –
What is a word, what is a sentence? problems of tokenization. *In : The 3rd International Conference on Computational Lexicography*, pp. 79–87. –
Budapest, 1994.


Schmid (Helmut). –
Probabilistic Part-of-Speech Tagging Using Decision Trees. *In : New Methods in Language Processing Studies in Computational Linguistics*, éd. par Jones (Daniel) et Somers (Harold).

Tsuruoka (Yoshimasa), Tateishi (Yuka), Kim (Jin-Dong), Ohta (Tomoko), McNaught (John), Ananiadou (Sophia) et Tsujii (Jun'ichi). –
Developing a Robust Part-of-Speech Tagger for Biomedical Text. *In : Proceedings of Advances in Informatics - 10th Panhellenic Conference on Informatics*, pp. 382–392.