APEC : de la sémantique dans les offres d'emploi

Retour au sommaire de la lettre

Une nouvelle série d'articles dans la lettre "Recherche & Référencement" avec la description de projets mis en place dans le cadre de la recherche d'information sur réseau intranet. Ce mois-ci, nous découvrons comment, pour faciliter l'intégration des CV et des offres d'emplois dans son système d'information, l'Apec utilise toutes les ressources offertes par les technologies du traitement de la langue...

Mettre en correspondance une demande portant sur un poste de consultant en "*business intelligence*" et un CV comportant des mots comme *informaticien* et *décisionnel* demande une expertise qui a longtemps été réservée à un traitement humain. Motivée par le nombre de documents qu'elle reçoit tous les mois (des milliers d'offres d'emploi), et encore plus de CV, l'Apec (agence pour l'emploi des cadres : http://www.apec.fr/) s'est lancée dans la mise en place d'outils de traitement de la langue permettant d'obtenir le même type de résultats.

L'agence a décidé d'ajouter une couche intelligente et automatisée dans le traitement des données textuelles internes et d'une partie du web. Dominique Jaquet, DSI de l'Apec, est convaincu de l'intérêt des outils de traitement automatique de la langue (TAL). "Une indexation sémantique améliore les résultats de recherche et aide à mesurer la qualité des documents" précise t-il.

Un traitement en trois étapes

Le projet global a été découpé en plusieurs "briques" :

- La première application mise en place analyse les offres afin d'en améliorer la qualité. Plus précisément, elle vérifie d'abord que les mentions légalement interdites (race, etc.) ne sont pas présentes, que toutes les mentions légales (rémunérations, expérience, lieux, etc.) sont, *a contrario*, bien présentes.

- Une deuxième phase nettoie l'annonce de tous les mots non significatifs et rattache chaque terme significatif (substantif ou expression) à un des champs devant figurer dans une annonce, une liste de champs définie par l'Apec. Par exemple, 'est rattaché au directeur' sera affectée au champ "dépendance hiérarchique".

- Une troisième étape consiste à évaluer la qualité globale de l'offre en fonction de son contenu. Cette dernière sera renvoyée à l'émetteur si trop de critères manquent ou sont mal formulés.

Le moteur d'indexation se fonde à la fois sur des statistiques, liées à la fréquence d'apparition des mots et expressions dans un texte, et sur la comparaison des mots avec le dictionnaire de 150 000 concepts fourni par l'éditeur Lingway. Avant d'affecter un sens à un mot, qui en possède trois en moyenne dans la langue française, le moteur mixe statistiques et proximité avec d'autres mots dans l'arborescence du dictionnaire.

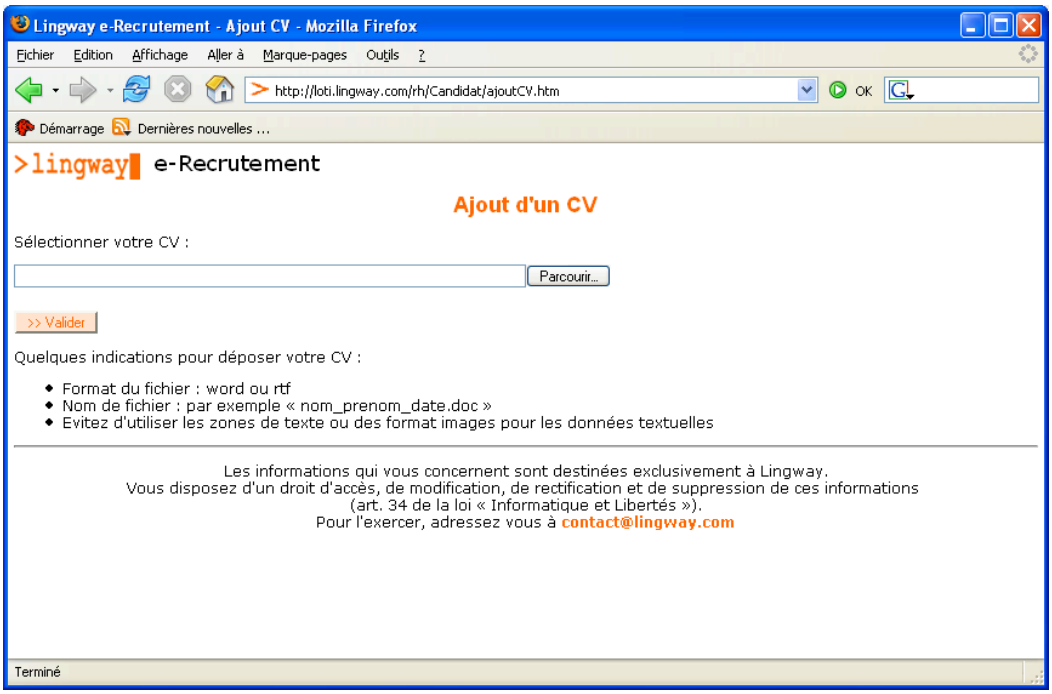
Organisé sous la forme d'une structure arborescente, ce dernier associe des synonymes à des concepts. Par exemple, la présence, dans la même annonce, des mots "système d'information" et "environnement" attribuent un sens technique, et non pas écologique.

Après ce premier projet mis en production l'année dernière, qui s'est étalé environ sur neuf mois, et a coûté 80 000 euros (divisés à parts égales entre la licence et le développement), l'Apec s'est lancée dans le traitement des CV. Première motivation du projet : faciliter la mise en correspondance des CV et des offres grâce au traitement sémantique. L'agence désirait également permettre l'intégration des CV au format bureautique dans ses bases de données sans imposer aux internautes de ressaisir leur CV sur son site.

Une nouvelle étape pour le traitement des CV

Baptisé e-Recrutement, le logiciel développé par Lingway initialement pour l'Apec extrait toutes les données d'un CV au format Word ou RTF pour les placer dans des champs structurés. "Cela répond aussi à une préoccupation des agences d'intérim qui estiment que les candidats remplissant les formulaires sur leur site sont souvent les moins intéressants et préfèrent ceux qui envoient leur CV sous forme de fichier bureautique", explique Hugues de Mazancourt, directeur technique de l'éditeur.

Cette brique n'évalue pas la qualité mais intègre tous les informations d'un CV dans une base de données. Après l'intégration dans le logiciel (*voir schéma ci-dessous*), elle suit schématiquement les étapes suivantes :



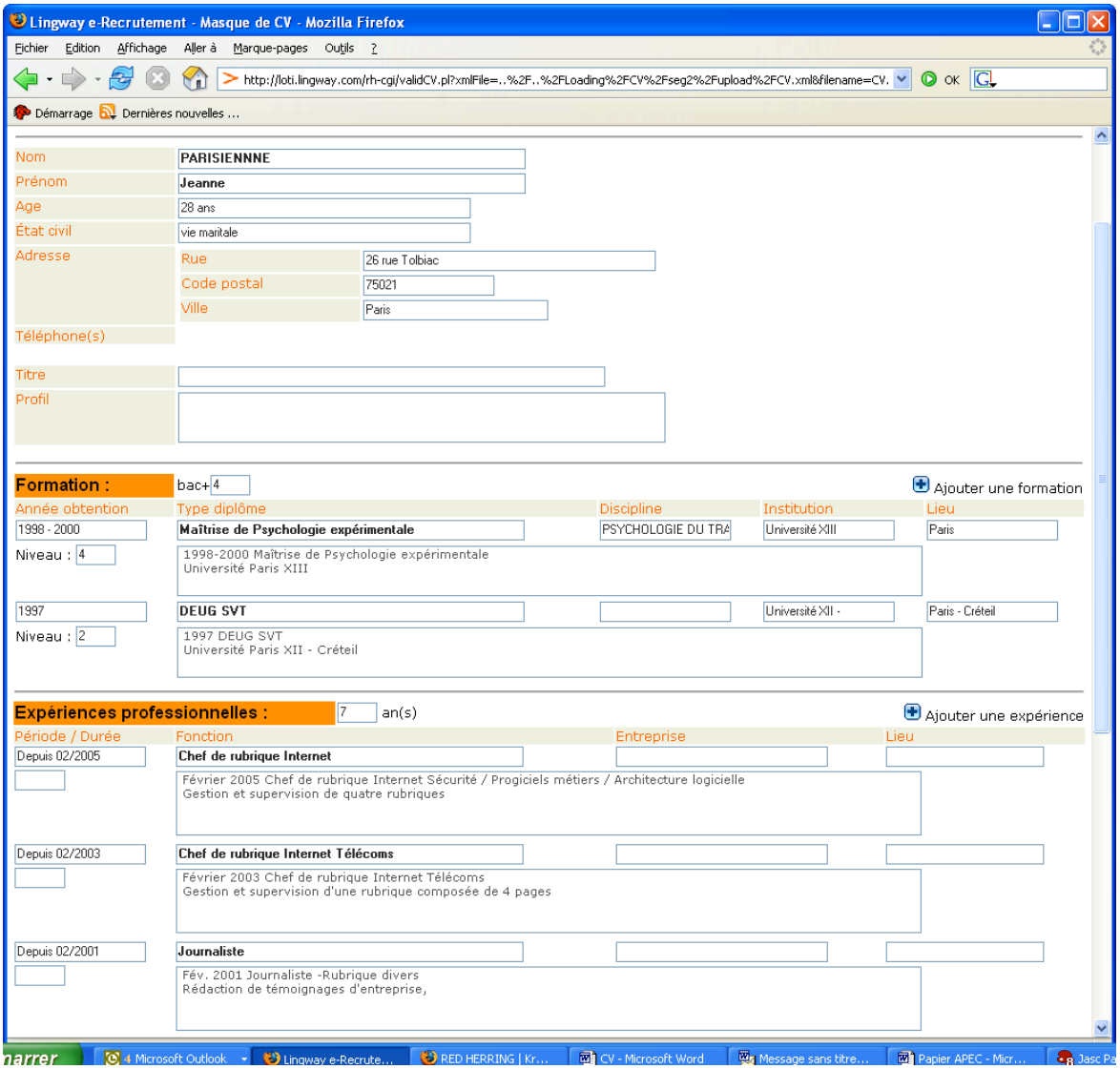
1. Reconnaissance de tous les mots avec identification des concepts (substantifs ou suite de substantifs) par rapport à des règles syntaxiques et par la comparaison avec les termes du thésaurus (dictionnaire organisé) fourni par l'éditeur.

2. Le moteur passe alors à la lemmatisation, une étape qui permet de ne retenir que le concept en ignorant les formes plurielles et conjuguées, des verbes notamment.

3. Grâce aux termes "métier" inclus dans le dictionnaire, les rubriques sont identifiées. Ce qui permet, par exemple, d'identifier les différentes façons de nommer l'expérience professionnelle (expériences, parcours, etc.).

4. Des règles "métier" attribuent un sens à un mot ou à une expression. Par exemple, "environnement" suivi d'un terme dans une liste de mots ("système", "Unix", etc.) attribue un sens technique à ce terme.

5. Les mots sont versés dans une base de données. Une interface de restitution permet de visualiser et si besoin modifier les données (*voir schéma ci dessous*).



Ce traitement offre plusieurs avantages. Outre l'automatisation de l'intégration des CV dans des bases de données, il permet de classer ces derniers dans plusieurs catégories. L'interrogation de la base de données passe ensuite par une phase d'interprétation sémantique. A partir d'une étape baptisée "expansion", le moteur interprétera l'expression "business intelligence" et la reliera à des termes comme "informaticien" et "décisionnel". L'interrogation des champs de la base de données se déroulera ensuite à partir de ces critères et renverra tous les CV ad hoc.

Conclusion

L'APEC va continuer à ajouter des applications sémantiques pour indexer d'autres documents comme des études économiques, des articles de presse, etc. D'ici un an ou deux, un candidat pourra paramétrer des alertes lui permettant de recevoir des offres d'emploi ou des articles économiques non pas à partir de simples critères SQL mais sur une interprétation du sens de sa question. Cette intégration des TAL sur l'intranet et le portail de l'agence améliore l'adéquation entre les recherches et les offres. Un "plus" indispensable pour que l'APEC puisse faire face à la concurrence montante des sites spécialisés dans ce domaine.

Patrick Brébion
Journaliste spécialisé