

Apache UIMA pour le Traitement Automatique des Langues

Nicolas Hernandez, Fabien Poulard, Stergos Afantenos, Matthieu Vernier et
Jérôme Rocheteau

LINA (CNRS - UMR 6241)

2 rue de la Houssinière – B.P. 92208, 44322 NANTES Cedex 3

{prenom.nom}@univ-nantes.fr

Résumé. L'objectif de la démonstration est d'une part de faire un retour d'expérience sur la solution logicielle Apache UIMA comme infrastructure de développement d'applications distribuées de TAL, et d'autre part de présenter les développements réalisés par l'équipe TALN du LINA pour permettre à la communauté de s'approprier ce « framework ».

Abstract. Our objectives are twofold : First, based on some common use cases, we will discuss the interest of using UIMA as a middleware solution for developing Natural Language Processing systems. Second, we will present various preprocessing tools we have developed in order to facilitate the access to the framework for the French community.

Mots-clés : Apache UIMA, Applications du TAL, Infrastructure logicielle.

Keywords: Apache UIMA, NLP applications, Middleware.

1 Contexte et objectifs

Plus que jamais, il est important que la communauté scientifique en Traitement Automatique des Langues (TAL) porte son attention sur les questions de conception et de développement d'applications de TAL et les enjeux d'ingénierie qui en découlent tels que la réutilisation, l'interopérabilité, le passage à l'échelle, la sérialisation (Leidner, 2003). Les motivations sont nombreuses : développer des applications distribuées de traitement d'informations semi-structurées multimédias et multilingues de plus en plus complexes, faciliter les échanges intra-équipe ainsi qu'entre partenaires académiques et industriels, capitaliser les savoir-faire et les produits développés, se donner la capacité de dépasser le cadre du prototypage pour offrir des solutions de force industrielle, gagner en réactivité face aux appels à projets nationaux et internationaux...

Dans le paysage des solutions logicielles existantes qui offrent des moyens d'intégration, de développement et de déploiement, le « framework » Apache UIMA¹ (*Unstructured Information Management Architecture*) constitue l'une des solutions les plus avancées et des plus prometteuses. Son objectif est de permettre l'utilisation et la construction d'applications distribuées visant l'analyse de contenus multimédias non structurés.

UIMA présente de nombreux atouts qui proviennent d'une part d'une dissociation très claire des considérations de l'utilisateur de chaînes de traitement, du développeur de composants métiers

¹incubator.apache.org/uima

et du développeur de l'infrastructure logicielle (« *middleware* ») des applications et du framework ; et d'autre part, des efforts de normalisation proposés pour la résolution de problèmes liés à ces différentes considérations.

Initié par IBM (Ferrucci & Lally, 2004), l'implémentation d'UIMA est aujourd'hui un projet en incubation au sein de l'ASF (*Apache Software Foundation*). Les principes de gestion de l'information non structurée (recherche sémantique et analyse de contenu) font l'objet d'un effort de standardisation de la part d'un comité technique de l'OASIS² (*Organization for the Advancement of Structured Information Standards*). UIMA bénéficie d'une communauté internationale active notamment dans le monde du logiciel libre. Sa distribution sous licence libre Apache n'interdit pas son utilisation au sein de produits commerciaux. Outre son framework pour la manipulation de données non structurées, ses outils (SDK (*Software Development Kit*)) développés par Apache pour faciliter son utilisation ou le développement de composants (notamment par le biais d'extensions pour l'IDE (*Integrated Development Environment*) Eclipse), UIMA apporte des solutions techniques pour les questions de composition de chaîne de traitement (« *workflow* »), de déploiement distribué (sur plusieurs processus ou sur des machines distantes), d'interopérabilité (par l'utilisation des standards pour la communication inter-composants par exemple via des services webs), de persistance possibles (stockage en XMI (*XML Metadata Interchange*), langage XML créé et standardisé par l'OMG³ (*Object Management Group*) pour l'échange de métadonnées UML)...

L'équipe TALN du LINA a beaucoup travaillé pour s'approprier cette architecture logicielle notamment en développant des composants connecteurs pour des outils extérieurs, en portant ses outils existants, en créant des paquets Debian du framework, en écrivant des tutoriaux, en organisant de formations internes pour les membres de l'équipe, en l'utilisant comme outil pédagogique dans les cours de Master et projets étudiants.

UIMA est désormais utilisé dans plusieurs de nos projets⁴ pour fédérer ou fournir des services de traitement à nos partenaires académiques et industriels : Dans l'ANR PIITHIE 2006-08 nous utilisons l'environnement UIMA pour développer et déployer des analyseurs sémantiques et discursifs comme services web de type REST de façon à détecter la réutilisation de textes. Dans l'ANR Blogoscopie 2006-08 nous développons un composant UIMA d'analyse d'opinions dans les blogs. Dans l'ANR C-Mantic 2007-09 nous avons pour objectif de développer un moteur de recherche sémantique avec UIMA pour la gestion des analyses sémantiques. Dans le projet régional Miles « Pays de Loire » 2007-09, nous utiliserons UIMA comme base architecturale pour connecter des composants distribués géographiquement dédiés à la reconnaissance de locuteurs dans des textes transcrits.

Cette démonstration rentre dans le cadre des travaux réalisés par l'équipe TALN du LINA pour la construction d'une communauté francophone⁵ de partenaires industriels et académiques notamment dans les domaines du traitement du langage naturel et de la parole autour de la solution logicielle Apache UIMA. Notre objectif est d'une part de faire un retour d'expérience sur la solution logicielle UIMA comme infrastructure de développement et d'échange, et d'autre part de présenter les développements réalisés (et en cours de réalisation) par l'équipe pour permettre à la communauté de s'approprier l'environnement.

²www.oasis-open.org/committees/uima

³www.omg.org

⁴www.piithie.com ; www.blogoscopie.org ; www.c-mantic.org

⁵www.uima-fr.org

2 Contenu de la démonstration

La démonstration portera sur trois axes. Tout d'abord nous effectuerons une présentation des outils de pré-traitement développés par le LINA. Ceux-ci couvrent les opérations de pré-traitement traditionnelles : reconnaissance du type MIME des fichiers manipulés, extraction des zones textuelles, reconnaissance de l'encodage d'un texte, de sa langue, segmentation en mots et phrases, analyse morphologique et syntaxique. Lors de ces travaux, nous avons cherché avant tout à réutiliser l'existant. Ainsi nous comptons parmi les composants développés des connecteurs pour les outils comme l'analyseur morphosyntaxique Tree-tagger de H. Schmid, l'étiqueteur grammatical de E. Brill, le lemmatiseur Flemm de F. Nammer, le projet Apache Lucene Tika. . .

Dans un deuxième axe, nous produirons un retour d'expérience sur l'utilisation de UIMA au travers de cas d'usage récurrents à nos projets de recherche tels que le déploiement en service web d'une chaîne de traitement, l'intégration d'outils hétérogènes, le déploiement sur plusieurs machines d'une application distribuée multi-composants... Ces utilisations seront illustrées à partir de systèmes que nous avons développés comme notre système de reconnaissance d'entités nommées, Nemesis (Fourour, 2002), et d'extraction terminologique, ACABIT (Daille, 2003).

Enfin, nous ouvrirons la discussion sur les questions d'interopérabilité en montrant comment des personnes désireuses d'adopter UIMA peuvent collaborer et échanger le résultat de leurs traitements (Hernandez *et al.*, ; Hahn *et al.*, 2007).

3 Moyens techniques demandés pour la présentation

Pour cette démonstration, nous aurions besoin à minima d'une table. Pour un meilleur rendu, nous apprécierions un vidéoprojecteur et son écran (ou bien un grand écran de PC), éventuellement une connexion Internet (cablée ou par wifi) et un support à poster.

Références

- DAILLE B. (2003). Conceptual structuring through term variations. In *F. Bond, A. Korhonen, D. MacCarthy and A. Villacencio (eds.), Proceedings ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, p. 9–16.
- FERRUCCI D. & LALLY A. (2004). Uima : an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4), 327–348.
- FOUOUR N. (2002). Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. In *Actes de TALN*, p. 265–274, Nancy.
- HAHN U., BUYKO E., TOMANEK K., PIAO S., MCNAUGHT J., TSURUOKA Y. & ANANIADOU S. (2007). An annotation type system for a data-driven nlp pipeline. In *LAW at ACL*, p. 33–40 : Prague, Czech Republic, June 28-29, 2007. Stroudsburg, PA : Association for Computational Linguistics.
- HERNANDEZ N., POULARD F., AFANTENOS S., VERNIER M. & ROCHETEAU J. Not yet another annotation framework but engineering-sensitive rules to design (uima) type systems. In *A paraître*.
- LEIDNER J. L. (2003). Current issues in software engineering for natural language processing. In *SEALTS at HLT/NAACL*, p. 45–50, Edmonton, Alberta, Canada.