86 :

Cet article présente une application permettant d'écrire des requêtes complexes sur des corpus étiquetés et de formater librement les résultats de ces requêtes. Le formalisme des requêtes est basé sur le principe des expressions régulières bien connu de la plupart des linguistes travaillant sur des corpus écrits. Contrairement à certains logiciels, qui ne permettent que l'extraction de concordances au format relativement figé, le formatage libre du résultat des requêtes permet leur réutilisation par des programmes ultérieurs et autorise une grande diversité d'applications, s'écartant largement du cadre des simples concordanciers.

205 :

Cet article expose la recherche effectuée dans le cadre de mon doctorat visant à élaborer un étiquetage morphologique de l'anglais et à désambiguïser automatiquement les ambiguïtés dues à la morphologie dans le cadre du projet LABELGRAM [9]. Nous montrons qu'il est très pertinent et efficace de travailler conjointement sur l'étiquetage et la désambiguïsation. Nous décrivons de manière précise notre contribution au système qui a consisté à mettre en place la partie anglaise. Pour ce faire, nous avons établi un dictionnaire en intention, nous avons évalué quantitativement le phénomène d'ambiguïté morphologique et établi la validité de la méthode de désambiguïsation par règles contextuelles pour l'anglais.

269 :

Cet article s'intéresse aux définitions formalisées de la base de données BDéf et montre en quoi la structure formelle de ces définitions est à même d'offrir une représentation originale de la polysémie lexicale.

352 :

Alors que de nombreux travaux portent actuellement sur la linguistique de corpus, l'utilisation de textes authentiques en classe de langue, ou de corpus dans l'enseignement des langues (via concordanciers), quasiment aucun travail n'a été réalisé en vue de la réalisation de bases de textes à l'usage des enseignants de langue, indexées en fonction de critères relevant de la problématique

de la didactique des langues. Dans le cadre de cet article, nous proposons de préciser cette notion d'indexation pédagogique, puis de présenter les principaux standards de description de ressources pédagogiques existants, avant de montrer l'inadéquation de ces standards à la description de textes dans l'optique de leur utilisation dans l'enseignement des langues. Enfin nous en aborderons les conséquences relativement à la réalisation de la base.

394 :

À travers la présentation de la plate-forme LinguaStream, nous présentons certains principes méthodologiques et différents modèles d'analyse pouvant permettre l'articulation de traitements sur corpus. Nous envisageons en particulier les besoins nés de perspectives émergentes en TAL telles que l'analyse du discours.

437 :

Dans cet article, nous cherchons à caractériser linguistiquement des segments textuels définis pragmatiquement, relativement à des besoins de réédition de documents et au sein desquels l'information est susceptible d'évoluer dans le temps. Sur la base d'un corpus de textes encyclopédiques en français, nous analysons la distribution de marqueurs textuels et discursifs et leur pertinence en nous focalisant principalement sur un traitement sémantique particulier de la temporalité.

477 :

Partant des lexiques TAL syntaxiques existants, cet article propose une représentation lexicale unifiée et normalisée, préalable et nécessaire à toute exploitation des lexiques syntaxiques hors de leur propre contexte de conception. Ce travail s'inscrit dans un cadre de modélisation privilégié ? le Lexical Markup Framework ? qui a été conçu dès le départ comme un modèle lexicographique intégrant les différents niveaux de description. Ce modèle permet d'articuler des descriptions extensionnelles et intensionnelles et fait référence à un jeu de descripteurs normalisés, garantissant la rigueur de la description des faits linguistiques et assurant, à terme, la compatibilité avec des formats de données utilisés pour l'annotation de corpus.

515 :

Nous présentons une approche empirique de l'évaluation automatique des réponses d'apprenants au sein d'un système d'Apprentissage des Langues Assisté par Ordinateur (ALAO). Nous proposons la mise en place d'un module d'analyse d'erreurs attestées sur corpus qui s'appuie sur des techniques robustes de Traitement Automatique des Langues (TAL). Cet article montre la réalisation d'un module d'analyse de morphologie flexionnelle, en situation hors-contexte, à partir d'un modèle linguistique existant.

1004 :

Dans le cadre d'apprentissages humains assistés par des environnements informatiques, les techniques de TAL ne sont que rarement employées ou restreintes à des tâches ou des domaines spécifiques comme l'ALAO (Apprentissage de la Langue Assisté par Ordinateur) où elles sont omniprésentes mais ne concernent que certaines dimensions du TAL. Nous cherchons à explorer les possibilités ou les performances des techniques voire des méthodes de TAL pour des systèmes moins spécifiques dès lors qu'une dimension de réseau et de collectivité est présente. Plus particulièrement, notre objectif est d'obtenir des indicateurs sur la construction collective de connaissances, et ses modalités. Ce papier présente la problématique de notre thèse, son contexte, nos motivations ainsi que nos premières réflexions.

1046 :

L'ouverture du Centre National de Réception des Appels d'Urgence (CNRAU) accessible aux sourds et malentendants fait émerger des questions linguistiques qui portent sur le français écrit des sourds, et des questions informatiques dans le domaine du traitement automatique du langage naturel. Le français écrit des sourds, pratiqué par une population hétérogène, comporte des spécificités morpho-syntaxiques et morpho-lexicales qui peuvent rendre problématique la communication écrite entre les personnes sourdes appelantes et les agents du CNRAU. Un premier corpus de français écrit sourd élicité avec mise en situation d'urgence (FAX-ESSU) a été recueilli dans la perspective de proposer des solutions TAL et linguistiques aux agents du CNRAU dans le

cadre de ces échanges écrits. Nous présentons une première étude lexicale, morpho-syntaxique et syntaxique de ce corpus reposant en partie sur une chaîne de traitement automatique, afin de valider les phénomènes linguistiques décrits dans la littérature et d'enrichir la connaissance du français écrit des sourds.

1209 :

Les « mots dièses» ou « hash tags » sont le moyen naturel de lier entre eux différents tweets. Certains « hash tags » sont en fait de petites phrases dont la décomposition peut se révéler particulièrement utile lors d'une analyse d'opinion des tweets. Nous allons montrer dans cet article comment l'on peut automatiser cette décomposition et cette analyse de façon à améliorer la détection de la polarité des tweets.

1220 :

La tâche du résumé multi-lingue vise à concevoir des systèmes de résumé très peu dépendants de la langue. L'approche par extraction est au coeur de ces systèmes, elle permet à l'aide de méthodes statistiques de sélectionner les phrases les plus pertinentes dans la limite de la taille du résumé. Dans cet article, nous proposons une approche de résumé multi-lingue, elle extrait les phrases dont les termes sont des plus discriminants. De plus, nous étudions l'impact des différents traitements linguistiques de base : le découpage en phrases, l'analyse lexicale, le filtrage des mots vides et la racinisation sur la couverture ainsi que la notation des phrases. Nous évaluons les performances de notre approche dans un contexte multi-lingue : l'anglais, l'arabe et le français en utilisant le jeu de données TAC MultiLing 2011.

1432 :

Nous proposons dans cet article une méthode semi-supervisée originale pour la création dereprésentations vectorielles pour des termes (complexes ou non) dans un espace sémantique pertinentpour une tâche de normalisation de termes désignant des entités dans un corpus. Notre méthodes'appuie en partie sur une approche de sémantique distributionnelle, celle-ci générant des vecteursinitiaux pour chacun des termes extraits. Ces vecteurs sont alors plongés dans un autre

espacevectoriel construit à partir de la structure d'une ontologie. Pour la construction de ce second espacevectoriel ontologique, plusieurs méthodes sont testées et comparées. Le plongement s'effectue parentraînement d'un modèle linéaire. Un calcul de distance (en utilisant la similarité cosinus) est enfineffectué pour déterminer la proximité entre vecteurs de termes et vecteurs de concepts de l'ontologieservant à la normalisation. La performance de cette méthode a atteint un rang honorable, ouvrantd'encourageantes perspectives.

1517 :

Un mésusage apparaît lorsqu'un patient ne respecte pas sa prescription et fait des actions pouvantmener à des effets nocifs. Bien que ces situations soient dangereuses, les patients ne signalentgénéralement pas les mésusages à leurs médecins. Il est donc nécessaire d'étudier d'autres sourcesd'information pour découvrir ce qui se passe en réalité. Nous proposons d'étudier les forums de santéen ligne. L'objectif de notre travail consiste à explorer les forums de santé avec des méthodes declassification supervisée afin d'identifier les messages contenant un mésusage de médicament. Notreméthode permet de détecter les mésusages avec une F-mesure allant jusqu'à 0,810. Cette méthodepeut aider dans la détection de mésusages et la construction d'un corpus exploitable par les expertspour étudier les types de mésusages commis par les patients.

1543 :

Les conversations techniques en ligne sont un type de productions linguistiques qui par de nombreuxaspects se démarquent des objets plus usuellement étudiés en traitement automatique des langues : ils'agit de dialogues écrits entre deux locuteurs qui servent de support à la résolution coopérative desproblèmes des usagers. Nous proposons de décrire ici ces conversations par un étiquetage en actes dedialogue spécifiquement conçu pour les conversations en ligne. Différents systèmes de prédictionsont été évalués ainsi qu'une méthode permettant de s'abstraire des spécificités lexicales du corpusd'apprentissage.