

Nom et Prénom : MOHAMED EL MAROUANI

Date de soutenance : 28/10/2021

Directeur de Thèse : ENNEYA NOURDDINE

Sujet de Thèse :

Machine Translation Quality and Recognizing Textual Entailment for Arabic: Evaluation and Error Analysis

Résumé :

Le Traitement Automatique de la Langue Naturelle (TALN) est un domaine de recherche composé de plusieurs tâches. Cette thèse s'intéresse à deux tâches qui sont l'Evaluation de la Traduction Automatique (ETA) et la Reconnaissance des Inférences Textuelles (RIT). Nous avons appliqué nos approches sur l'Arabe Standard Moderne. Dans la première partie, nous exposons nos contributions au développement de l'ETA pour l'arabe. Nous présentons la métrique AL-TERp dédiée pour l'arabe et qui intègre la connaissance linguistique pour couvrir plusieurs aspects relatifs aux phrases Hypothèse et Référence. AL-TERp réalise des résultats prometteurs en termes de corrélation avec les jugements humains. En outre, nous avons présenté dans la deuxième partie l'analyse statistique des erreurs de la traduction automatique. Cette analyse intègre pour la première fois les Modèles des Liens Cumulatifs pour examiner la relation entre les erreurs obtenues automatiquement et le jugement des annotateurs humains. La dernière partie est consacrée en premier lieu pour la description de nos deux systèmes de RIT en arabe. Ces systèmes sont basés sur l'apprentissage automatique et les techniques d'optimisation et intègrent un composant important d'étiquetage et de normalisation des expressions temporelles. Pareillement à l'ETA, nous avons amélioré la performance des systèmes de RIT via l'intégration des connaissances linguistiques à partir de plusieurs ressources. Ces systèmes atteignent des résultats très encourageants. Pour investiguer l'intégration fonctionnelle des deux tâches, nous présentons aussi une métrique ETA basée essentiellement sur le système de RIT. Les expérimentations effectuées ont montré l'importance de l'intégration mutuelle des tâches du TALN.

Abstract :

Natural Language Processing (NLP) is a research field divided into several tasks. In this thesis, we have dealt with two important tasks namely Machine Translation Evaluation (MTE) and Recognizing Textual Entailment (RTE). We have focused on Modern Standard Arabic language to test our proposed approaches. In the first part, we expose our contribution to the improvement of MTE applied to Arabic. We present an extended MTE metric, AL-TERp, dedicated to this language and aiming to integrate linguistic knowledge to capture several aspects related to Hypothesis and Reference sentences. AL-TERp achieves promising results in terms of correlation with human judgments. To continue our study of machine translation quality, we dedicated the second part to the statistical analysis of machine translation errors. This study involves for the first time the Cumulative Link Models in the analysis of the relation between occurred errors and the human judgment of machine translation outputs. The last part firstly presents our two RTE systems devoted to Arabic. These two systems are based on machine learning and optimization techniques and integrate an important component of temporal tagging. Similarly to the MTE system, we improved the performance of RTE systems via the integration of linguistic knowledge from several resources. These systems achieve encouraging results. To employ the previous results and to explore the combination of the two NLP tasks, we present in the last part a new MTE metric based essentially on our RTE engine. Performed experiments have demonstrated the importance of mutual integration of natural language understanding tasks.