

Intégration de méthodes de TAL à des méthodes d'apprentissage pour l'évaluation de réponses à des questionnaires

Encadrantes : Brigitte Grau (PU, co-directeur, LIMSI) et et Ludovic Denoyer (MCF-HDR, co-directeur, LIP6) et Anne-Laure Ligozat (LIMSI) et Benjamin Piwowarski (LIP6) co-encadrants

Contexte

Le contexte général de la thèse porte sur la compréhension de texte, et plus particulièrement sur la mise en place de méthodes capables d'évaluer si les réponses données à des questions sont correctes ou non.

Cette problématique constitue un champ de recherche assez récent comme le montre la création de campagnes d'évaluation autour de ce thème, QA4MRE¹ et la tâche 7 de Semeval² : The Joint Student Response Analysis.

Afin d'évaluer le niveau de compréhension d'un être humain, un élève par exemple, un mode couramment utilisé est la production de questionnaires auxquels l'élève devra répondre. Ces tests peuvent avoir la forme de QCM posé sur un texte ou un domaine ou de questionnaires où les réponses doivent être rédigées. Ce mode d'évaluation peut être réutilisé pour évaluer la compréhension que peut développer une machine. Dans le cas de QCM, la machine a pour but de remplir le QCM (elle choisit la bonne réponse). Dans le second cas, la machine évalue la réponse donnée par l'élève par rapport à la correction donnée par l'enseignant.

Cette problématique peut être posée comme un problème d'implication textuelle : étant donné un texte T et une hypothèse H, déterminer si $T \Rightarrow H$, c'est à dire, est-ce que l'on peut raisonnablement déduire l'information portée par l'hypothèse à partir des informations contenues dans le texte et dans une base de connaissance d'arrière plan.

Dans le cas de QCM, les hypothèses sont formées à partir de la question et de chaque réponse candidate, et le texte considéré est le texte sur lequel sont posées les questions, ou plus souvent un passage de texte que l'on sélectionne préalablement. Dans le cas de questionnaires ouverts, le texte est formé par la correction de référence et l'hypothèse est la réponse à évaluer, dans le contexte de la question posée.

Cette thèse s'inscrit dans un cadre applicatif de e-learning qui consiste à développer des outils pour permettre à des étudiants de s'auto-évaluer et à des enseignants de disposer d'aide à l'évaluation des étudiants et leur suivi. Cet environnement permettra de disposer de corpus différents selon les niveaux des élèves et les domaines enseignés. Nous partirons des corpus existants fournis par les évaluations, en anglais, mais nous veillerons à proposer des solutions aussi indépendantes que possible de la langue, ou, du moins, adaptables au français.

Sujet

Le problème posé par l'implication textuelle est de reconnaître différentes formulations d'une même information et les relations d'inférence entre elles. Il s'agit donc de tenir compte des variations de termes et du type d'expression des énoncés pour s'affranchir de la forme de surface. Cela peut être abordé par l'utilisation de ressources lexicales et en prenant en compte les relations syntaxiques voire sémantiques que possèdent les termes dans les phrases et les relations discursives entre énoncés. Il est également nécessaire d'utiliser des connaissances d'arrière plan – textes du même domaine ou de domaine général – pour exploiter les connaissances qu'ils contiennent (Penas et al., 2011).

Une approche développée par le LIMSI est fondée sur l'utilisation de ressources linguistiques et la

1 <http://celct.fbk.eu/QA4MRE/>

2 <http://www.cs.york.ac.uk/semeval-2013/task7/>

structure syntaxique des phrases pour comparer deux énoncés en adaptant des méthodes de calcul de distance de la littérature.

Le LIP6 s'intéresse au problème de l'apprentissage de représentations (Bengio et al., 2012). Le but est de s'affranchir d'un travail manuel de pré-traitement des données pour représenter de manière automatique et semi-supervisée des objets complexes tels que des parties de réseaux sociaux ou des textes (Socher et al., 2011). Ce domaine est en pleine émergence comme en témoigne l'organisation de la première conférence³ en 2013. Basé sur ce type de technique, il est par exemple possible d'apprendre à classifier la relation de causalité entre deux parties d'un texte.

L'objectif de cette thèse est d'étudier comment intégrer ces approches dans un cadre unifié dans le but de développer des techniques novatrices de représentation/compréhension du texte, en profitant de l'expérience complémentaire du LIMSI (traitement du langage naturel et tâche de question/réponse), et sur celle du LIP6 (apprentissage automatique sur des données complexes). En particulier, cette thèse abordera les points suivants :

- produire et faire varier les niveaux de représentation des passages : suite de termes, structure syntaxique, structure sémantique, structure discursive ;
- étudier comment et pour lesquels de ces objets un apprentissage automatique des représentations permet d'améliorer les performances, et en particulier exploiter des techniques d'apprentissage automatique afin d'inférer des relations entre des passages ;
- étudier en quoi la variation des niveaux de représentations données en entrée peut amener à des processus ou résultats d'apprentissage différents.

L'évaluation des solutions proposées sera faite en premier lieu par la participation aux campagnes d'évaluation existantes. Leur intégration dans l'environnement de e-learning permettra leur évaluation par des utilisateurs.

Références :

Bengio, Y., Courville, A., & Vincent, P. (2012). Representation Learning: A Review and New Perspectives. ArXiv 1206.5538, 2012.

K. D. Forbus, C. Riesbeck, L. Birnbaum, K. Livingston, A. Sharma, L. Ureel , Integrating Natural Language, Knowledge Representation and Reasoning, and Analogical Processing to Learn by Reading , Conference of the Association for the Advancement of Artificial Intelligence , 2007

Mulkar R., Hobbs J.R. , and Hovy E.H., *Learning from Reading Syntactically Complex Biology Texts*, Proceedings of the AAI Spring Symposium Commonsense'07. Stanford University, CA, 2007

Penas A. et al, 2011, *Overview of QA4MRE at CLEF 2011: Question Answering for Machine Reading Evaluation*, Conference CLEF, <http://celct.fbk.eu/QA4MRE/>

Penas A. et Hovy E.H., Filling Knowledge Gaps in Text for Machine Reading , Coling 2010: Poster Volume, pages 979–987, Beijing, August 2010

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., & Manning, C. D. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Advances in neural information processing systems*, 24, 801–809.

3 First International Conference on Learning Representations (ICLR 2013)