Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Rachel BAWDEN : rachel.bawden@ed.ac.uk

Titre : Au-delà de la phrase : traduction automatique de dialogue en contexte

Mots-clés : Traduction automatique, contexte, dialogue, discours.

Title: *Going beyond the Sentence: Contextual Machine Translation of Dialogue*

Keywords: *Machine translation, context, dialogue, discourse.*

Thèse de doctorat en informatique, LIMSI, CNRS, Université Paris-Sud, Université Paris-Saclay, Orsay, sous la direction de Sophie Rosset (DR, CNRS, LIMSI) et Thomas Lavergne (MC, Université Paris-Sud, Université Paris-Saclay, LIMSI). Thèse soutenue le 29/11/2018.

Jury : Mme Sophie Rosset (DR, CNRS, LIMSI, codirectrice), M. Thomas Lavergne (MC, Université Paris-Sud, Université Paris-Saclay, LIMSI, codirecteur), M. Nicolas Sabouret (Pr, Université Paris-Sud, Université Paris-Saclay, LIMSI, président), M. Jörg Tiedemann (Pr, Université d'Helsinki, Finlande, rapporteur), M. Loïc Barrault (MC, Université du Mans, rapporteur), Mme Lucia Specia (Pr, Université de Sheffield et Imperial College London, Royaume-Uni, examinatrice), M. Andrei Popescu-Belis (Pr, Haute École d'Ingénierie et de Gestion du Canton de Vaud, Suisse, rapporteur).

Résumé : *Les systèmes de traduction automatique (TA) ont fait des progrès considérables ces dernières années. La majorité d'entre eux reposent pourtant sur l'hypothèse que les phrases peuvent être traduites indépendamment les unes des autres. Ces modèles de traduction ne s'appuient que sur les informations contenues dans la phrase à traduire. Ils n'ont accès ni aux informations présentes dans les phrases environnantes ni aux informations que pourrait fournir le contexte dans lequel ces phrases ont été produites.*

La TA contextuelle a pour objectif de dépasser cette limitation en explorant différentes méthodes d'intégration du contexte extraphrastique dans le processus de traduction. Les phrases environnantes (contexte linguistique) et le contexte de production des énoncés (contexte extralinguistique) peuvent fournir des informations cruciales pour la traduction, notamment pour la prise en compte des phénomènes discursifs et des mécanismes référentiels.

La prise en compte du contexte est toutefois un défi pour la traduction automatique. Évaluer la capacité de telles stratégies à prendre réellement en compte le contexte et à améliorer ainsi la qualité de la traduction est également un problème délicat, les métriques d'évaluation usuelles étant pour cela inadaptées, voire trompeuses.

Dans cette thèse, nous proposons plusieurs stratégies pour intégrer le contexte, tant linguistique qu'extralinguistique, dans le processus de traduction. Nos expériences s'appuient sur des méthodes d'évaluation et des jeux de données que nous avons développés spécifiquement à cette fin. Nous explorons différents types de stratégies : les stratégies par prétraitement, où l'on utilise le contexte pour désambiguïser les données fournies en entrée aux modèles ; les stratégies par post-traitement, où l'on utilise le contexte pour modifier la sortie d'un modèle non contextuel, et les stratégies où l'on exploite le contexte pendant la traduction proprement dite. Nous nous penchons sur de multiples phénomènes contextuels, et notamment sur la traduction des pronoms anaphoriques, la désambiguïsation lexicale, la cohésion lexicale et l'adaptation à des informations extralinguistiques telles que l'âge ou le genre du locuteur. Nos expériences, qui relèvent pour certaines de la TA statistique et pour d'autres de la TA neuronale, concernent principalement la traduction de l'anglais vers le français, avec un intérêt particulier pour la traduction de dialogues spontanés.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-02004683

Matthieu LABEAU : mlabeau@exseed.ed.ac.uk

Titre : Modèles de langue neuronaux : gestion des grands vocabulaires

Mots-clés : Réseaux de neurones, modèles de langue, grands vocabulaires.

Title: *Neural Language Models: Dealing with Large Vocabularies*

Keywords: *Neural networks, language models, large vocabularies.*

Thèse de doctorat en informatique, LIMSI, CNRS, Université Paris-Sud, Université Paris-Saclay, Orsay, sous la direction de Alexandre Allauzen (Pr, Université Paris-Sud, Université Paris-Saclay, LIMSI). Thèse soutenue le 21/09/2018.

Jury : M. Alexandre Allauzen (Pr, Université Paris-Sud, Université Paris-Saclay, LIMSI, directeur), M. Pierre Zweigenbaum (DR, CNRS, LIMSI, président), M. Massih-Reza Amini (Pr, Université Grenoble-Alpes, rapporteur), M. Phil Blun-

som (associate professor, University of Oxford, Royaume-Uni, rapporteur), M. Armand Joulin (research scientist, Facebook Artificial Intelligence Research, examinateur), M. André Martins (research scientist, Instituto de Telecomunicaçoes, examinateur).

Résumé : *Le travail présenté dans cette thèse explore les méthodes pratiques utilisées pour faciliter l'entraînement et améliorer les performances des modèles de langues munis de très grands vocabulaires. La principale limite à l'utilisation des modèles de langue neuronaux est leur coût computationnel : il dépend de la taille du vocabulaire avec laquelle il grandit linéairement. La façon la plus aisée de réduire le temps de calcul de ces modèles reste de limiter la taille du vocabulaire, ce qui est loin d'être satisfaisant pour de nombreuses tâches. La plupart des méthodes existantes pour l'entraînement de ces modèles à grand vocabulaire évitent le calcul de la fonction de partition, qui est utilisée pour forcer la distribution de sortie du modèle à être normalisée en une distribution de probabilités. Ici, nous nous concentrons sur les méthodes à base d'échantillonnage, dont l'échantillonnage par importance et l'estimation contrastive bruitée. Ces méthodes permettent de calculer rapidement une approximation de cette fonction de partition. Cependant, elles varient en efficacité, surtout dans le cas des très grands vocabulaires.*

L'examen des mécanismes de l'estimation contrastive bruitée nous permet de proposer des solutions qui vont considérablement faciliter l'entraînement, ce que nous montrons expérimentalement. Nous discutons notamment de l'impact de la distribution de bruit choisie pour échantillonner, ainsi que des autres hyperparamètres impliqués dans l'apprentissage. Ensuite, nous utilisons la généralisation d'un ensemble d'objectifs basés sur l'échantillonnage en tant que divergences de Bregman pour expérimenter avec de nouvelles fonctions objectifs. Enfin, nous exploitons les informations données par les unités sous-mots (caractères ou décompositions morphologiques) pour enrichir les représentations de mots.

À l'aide des méthodes d'apprentissage présentées dans la partie précédente, nous cherchons à entraîner des modèles munis de ces représentations, pour les mots en entrée et surtout en sortie du modèle. Nous expérimentons avec différentes architectures sur le tchèque et nous montrons que les représentations basées sur les caractères permettent l'amélioration des résultats pour les grands vocabulaires, d'autant plus lorsque l'on réduit conjointement l'utilisation des représentations des mots les plus rares, qu'il est difficile d'apprendre.

URL où le mémoire peut être téléchargé :
http://www.theses.fr/2018SACLS313

Caroline LANGLET : langlet.caro@gmail.com

Titre : Analyse des sentiments dans les conversations humain-agent. Vers un modèle des goûts de l'utilisateur

Mots-clés : Interaction humain-agent, agents conversationnels animés, analyse de sentiments.

Title: *Sentiment Analysis in Human-Agent Conversations. Modelling User's Likes and Dislikes*

Keywords: *Human-agent interaction, embodied conversational agent, sentiment analysis.*

Thèse de doctorat en informatique, LTCI (Laboratoire Traitement et Communication de l'Information), IDS (Image, Données, Signal), Télécom ParisTech, Paris, sous la direction de Catherine Pelachaud (DR, CNRS, ISIR) et Chloé Clavel (MC, Télécom Paristech). Thèse soutenue le 26/09/2018.

Jury : Mme Catherine Pelachaud (DR, CNRS, ISIR, codirectrice), Mme Chloé Clavel (MC, Télécom Paristech, codirectrice), M. Bjorn Schuller (Pr, Imperial College of London, Royaume-Uni, rapporteur), M. Thierry Poibeau (DR, CNRS, Lattice, rapporteur), Mme Pascale Sébillot (Pr, INSA de Rennes, IRISA, présidente), M. Dirk Heylen (Pr, Université de Twente, Pays-Bas, examinateur), Mme Marie-Jeanne Lésot (MC, Sorbonne Université, examinatrice), M. Nicolas Maudet (Pr, Sorbonne Université, examinateur).

Résumé : *Cette thèse se situe à la croisée de deux domaines de recherche : celui du sentiment analysis et celui des agents conversationnels animés. Les agents conversationnels animés peuvent être définis comme des personnages virtuels ayant la capacité de converser avec un utilisateur humain. Afin d'accroître les compétences communicationnelles de l'agent, il est important que celui-ci soit doté d'une forme d'intelligence socio-émotionnelle. L'agent doit être ainsi en capacité de gérer des signaux socio-émotionnels, tant du côté de la génération que de celui de la détection.*

Du côté de la génération, de nombreux travaux ont produit des modèles optimisant la production de gestes ou d'expressions faciales pour exprimer soit des émotions soit des attitudes sociales. Du côté de la détection, une majorité des travaux se concentrent sur l'analyse d'indices socio-affectifs non verbaux (expressions faciales, indices acoustiques). Le contenu verbal et les expressions de sentiment qu'il véhicule restent quant à eux encore partiellement exploités. En effet, les rares études intégrant un module de détection des sentiments de l'utilisateur dans le cadre de conversations humain-agent ne prennent pas en compte les spécificités de ce contexte d'interaction.

Pour combler cette lacune, notre travail s'intéresse à l'analyse du contenu verbal produit par l'utilisateur et à la manière dont celui-ci réfère à ou exprime des sentiments, des affects ou des attitudes. Nous en proposons un modèle de détection au cours d'une interaction multimodale et en face à face avec un agent conversationnel animé. Pour

construire ce modèle, deux questions se sont posées à nous. Dans un premier temps, il nous a fallu identifier, au sein de la vaste classe des expressions de sentiment, celles qui apparaissaient comme les plus pertinentes pour l'élaboration des stratégies de communication de l'agent. Dans un second temps, nous avons dû choisir une méthode devant être non seulement opérante pour une analyse à grain fin de ces expressions, mais également adaptable au contexte conversationnel.

Nos contributions s'articulent autour de trois axes. Tout d'abord, nous fournissons un modèle linguistique des expressions de sentiment dans une conversation humain-agent. Trois unités conversationnelles sont considérées : le tour de parole, la paire adjacente et la séquence thématique. Cette analyse met en évidence un certain nombre de caractéristiques nécessaires au développement d'un ensemble de règles de détection. Ensuite, nous proposons un modèle de détection symbolique intégrant des règles sémantiques et des grammaires formelles. Ce modèle repose sur une analyse ascendante des énoncés — du niveau lexical au niveau phrastique — et se concentre successivement sur trois cadres d'analyse : le tour de parole, la paire adjacente et la séquence thématique. Enfin, nous proposons un protocole d'évaluation pour la validation des règles. Grâce à la création de deux plateformes d'annotation, nous avons pu créer deux jeux d'annotations sur deux corpus différents : un corpus de type small-talk *et un corpus de négociation. Les performances du système ont ainsi pu être évaluées par rapport aux références obtenues.*

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-02002580

Elvys LINHARES PONTES : elvyslpontes@gmail.com

Titre : Résumé translingue de textes par compression

Mots-clés : Résumé translingue de textes, compression multi-phrases, multilinguisme, optimisation.

Title: *Compressive Cross-Language Text Summarization*

Keywords: *Cross-language text summarization, multi-Sentence compression, multilingual, optimization.*

Thèse de doctorat en informatique, Laboratoire Informatique d'Avignon (LIA), Centre d'Enseignement et de Recherche en Informatique (CERI), Avignon Université, Avignon, sous la direction de Torres-Moreno Juan-Manuel (MC HDR, Avignon Université, LIA). Thèse soutenue le 30/11/2018.

Jury : M. Torres-Moreno Juan-Manuel (MC HDR, Avignon Université, LIA, codirecteur), M. Stéphane Huet (MC, Avignon Université, LIA, codirecteur), Mme Andréa Carneiro Linhares (MC, Universidade Federal do Ceará, Brésil, codirectrice), Mme Marie-Francine Moens (Pr, KU Leuven, LIIR, Louvain, Belgique, rapporteur),

M. Antoine Doucet (Pr, Université de La Rochelle, L3i, rapporteur), M. Frédéric Béchet (Pr, Aix Marseille Université, LIS, président), M. Guy Lapalme (Pr, Université de Montréal, DIRO, Canada, examinateur), Mme Fatiha Sadat (Pr, Université du Québec à Montréal, GDAC-LIA, Canada, examinatrice), M. Petko Valtchev (Pr, Université du Québec à Montréal, GDAC-LIA, Canada, examinateur), M. Florian Boudin (MC, Université de Nantes, LS2N, examinateur).

Résumé : *The popularization of social networks and digital documents has caused a rapid increase of the information available on the Internet. However, this huge amount of data cannot be handled manually. Natural Language Processing (NLP) deals with interactions between computers and human languages in order to process and analyze natural language data. NLP techniques incorporate a variety of methods, including linguistics, statistics or machine learning, to extract entities, relationships or understand a document. In this thesis, among several existing NLP applications, we are interested in cross-language text summarization which produces a summary in a language different from the language of the source documents. We also look at other NLP tasks (word encoding representation, semantic similarity, sentence and multi-sentence compression) to generate more stable and informative cross-lingual summaries.*

Most NLP applications, including text summarization, rely on a similarity measure to analyze and to compare the meaning of words, chunks, sentences and texts. A way to analyze similarity is to generate a representation for sentences that takes into account their sense. The meaning of sentences is defined by several elements, such as the context of words and expressions, word order and previous information. Simple metrics, such as cosine metric and Euclidean distance, provide a measure of similarity between two sentences. However, they put aside the order of words or multi-words. To overcome these limitations, we propose a neural network model that combines recurrent and convolutional neural networks to estimate the semantic similarity of a pair of sentences (or texts) from both the local and general contexts of words. On a supervised task, our model predicts more accurate similarity scores than baselines by taking greater account of the local and the general meanings of not only words, but also multi-word expressions.

In order to remove redundancies and non-relevant information of similar sentences, we propose a multi-sentence compression method that abbreviates and fuses them in a correct and short sentence that contains the main information. First, we model clusters of similar sentences as word graphs. Then, we apply an integer linear programming model that guides the compression of these clusters based on a list of keywords. We look for a path in the word graph that has a good cohesion and contains the maximum of keywords. Through a series of experiments, we show that our approach outperforms baselines by generating more informative and correct compressions for French, Portuguese and Spanish languages.

Finally, we combine these previous methods to build a cross-language text summarization system. Our system is an {English, French, Portuguese, Spanish}-to-{English, French} cross-language text summarization framework that examines the information in source and target languages to identify the most relevant sentences. Inspired

by the compressive text summarization studies in monolingual analysis, we adapt our multi-sentence compression method for this problem to just keep the main information. Our system proves to be a good alternative to compress redundant parts and to preserve relevant information, without losing grammatical quality. Experimental analysis of {English, French, Portuguese, Spanish}-to-{English, French} cross-lingual summaries indicate that our approach significantly outperforms the state of the art for all these languages. Besides, we apply cross-language summarization and discuss its role in two applications: microblog contextualization and spech-to-text summarization. In the last case, our method still achieves better and more stable scores, even for transcript documents that have grammatical errors and missing information.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-02003886