Projet de collaboration avec *(FAP, BRESIL)*
(2014-2016)
Responsable scientifique : *Laurent Besacier,LIG – UMR 5217, INSII*

RAPPORT D'ACTIVITÉ 2014

A. MOBILITE TRANSNATIONALE

A.1 - Accueil, dans le laboratoire français, de chercheurs du laboratoire partenaire étranger
Objet de l'accueil, date, nom du chercheur, qualité, laboratoires d'origine et d'accueil, durée du séjour, si le chercheur a donné un séminaire indiquer le titre

Aucun (nous avons été notifiés mi-2014 ce qui n'a pas permis d'organiser suffisamment à l'avance l'accueil de chercheurs brésiliens en France sur ce projet)

A.2 - Séjours, dans le laboratoire partenaire étranger, de chercheurs du laboratoire français
Objet du séjour, date, nom du chercheur, qualité, laboratoires d'origine et d'accueil, durée du séjour, si le chercheur a donné un séminaire indiquer le titre

3 séjours de chercheurs français ont été financés pour une mission à San Carlos (Brésil) du 27 Octobre au 3 Novembre 2014. Les trois participants ont été :

- Laurent BESACIER (Professeur, LIG)
- Agnès TUTIN (Professeur, LIG)
- Isabelle TELLIER (Professeur, LATTICE)

Un compte-rendu de ce séjour et les présentations réalisées sont disponibles en ligne sur :
http://aim-west.imag.fr/program-aim-west-workshop-29-31-oct-2014-san-carlos-brasil/

A.3 – Organisation de réunions de travail, séminaires etc. dans le cadre du projet
Indiquer l'objet de la réunion, date, lieu (laboratoire partenaire ou ville), nombre total de participants, identifier les participants français (nom, qualité, laboratoire de rattachement, durée de la mission). Vous pouvez également donner ces renseignements sous forme de tableau Excel.

Un bilan des réunions de travail sur le projet est donné ci-dessous (en anglais). Il se trouve aussi en ligne sur le site Web du projet AIM-WEST (http://aim-west.imag.fr/project-meetings/)

- **March 26, 2014 – Skype** – LATTICE, UFRGS, UFScar, LIG, LIF (Thierry P. Aline V. Helena C. Laurent B. Carlos R.) -> Kick-off meeting
- **April 16, 2014 – Skype** – LATTICE, UFRGS, LIG, LIF (Viviane M. Aline V. Laurent B. Frédéric B. Alexis N. Benoit F.) -> November missions/meetings planning
- **May 22, 2014 – ENS Paris (France)** – LATTICE, UFRGS, LIG, LIF (Thierry P. Aline V. Marco I. Laurent B. Carlos R.) -> Scientific goals and internship subjects
- **29-31 October 2014** AIM-WEST meeting was held in **San-Carlos (Brasil)**
- **January 21st, 2015** – Short **Skype** meeting between AIM-WEST partners

B. TRAVAUX EN COLLABORATION

B.1 – Etat d'avancement du projet scientifique
5 pages maximum pour l'année en cours ou 10 pages maximum pour les projets arrivant à terme. Le nom des chercheurs impliqués sera précisé

(en anglais)
The project has just started and the following workplan has been decided (mostly during the AIM-WEST workshop held in Sao Carlos this year) :

Task 1 Release a usable Fr-En-Pt-BR corpus : could be extracted from TED , then useful to train first SMT systems, include them in LALIC portal => Summer Internship from Polytech' Grenoble (RICM4) co-supervised by Helena & Laurent on this topic

Task 2 Organize internally a Fr-En MT shared task : goal evaluate the behavior of several MT systems on translating several types of MWEs
a) Choose a corpus (potentially annotated in MWEs) : we can take the French corpus annotated in Grenoble – it has 548 sentences – 700 MWEs
(might be good to separate sentences by MWE categories – if there are multiple MWEs in a single sentence, then we duplicate it)

b) Translate it with several systems (LIF, LIG, Systran, Google)

c) Define metrics (auto. ; man.) & evaluate – use Sectra_W for PE plateform for instance
(En is the target language so we can use TERpa and METEOR stuff…)
auto. eval: LIG try to come up with something
post-edition: get posteditions and then back to automatic evaluation metric but using PEditions as reference
manual: Agnes thinks we need it ;

Possible Schedule :
-January 2015 – Train/Test data released
-30th March 2015 – Collecting MT systems output
-April-May 2015 – Postediting system outputs (Sectra_W && discuss with Lingxiao)
-AIM-WEST workshop in October 2015 (Porto-Alegre)

Task 3 extend MWE annotation done in French to English (Carlos+Agnès) and Portuguese (Helena's group) using similar typology (Agnès will provide the Eng typology + exemples of expressions in several languages)

Task 4 POS tagging or morphological analysis of non contiguous MWEs

-extend Tellier & Constant's work to non contiguous MWEs
-Christian & Paltonio work on handling non contiguouas MWE on morphological analysis of portuguese

Speech Processing aspects are reported to 2016 !! (to prepare this, start thinking of taken into account lattices in MWE toolkit?)

Task 5 Using word alignments to detect MWEs in parallel corpora

The idea is to propose and internship to extend the work Helena, Aline and Carlos started in 2010 (LRE special issue paper). We would like to use automatic word alignment to detect MWEs automatically and integrate it in MWEtoolkit. A detailed internship subject will be posted online soon.

B.2 – Autres activités communes
Activités avec des chercheurs du laboratoire partenaire étranger hors du contexte du projet, projets co-déposés dans le cadre d'appels nationaux ou européens, contrats industriels,…
Objet, cadre, dates, bref descriptif.

Un certain nombre de stages co-encadrés france-brésil ont été proposés et sont en ligne sur : http://aim-west.imag.fr/research-internships/

Ils ont également donnés ci-dessous pour information.

1) Statistical Machine Translation (SMT) of Subtitles for English, French and Portuguese Languages and better handling of Multi Word Expressions in SMT

Supervisors : L. Besacier & H. Caseli (Laurent.BesacierATimag.fr – helenacaseliATdc.ufscar.br)

This intership is done in the context of a larger project which aims at proposing innovative and efficient methods for handling Multi-Word-Expressions (MWEs) in Statistical Machine Translation (SMT). A corpus for specifically evaluating this aspect will first have to be collected for SMT between English, French anf Portuguese languages.

This corpus could be extracted from TED corpus (of subtitles – https://wit3.fbk.eu/) and first machine translation systems will be built using Moses toolkit. After that, some research work to provide better handling of Multi Word Expressions in SMT will follow. Some ideas around this project are the following : work on data pre-processing to improve word-alignment and translation models, propose sparse features (related to MWEs) in the SMT model, work on N-best list re-ranking or graph re-decoding, etc.

2) Extending the MWEtoolkit with token identification – TAKEN

Supervisors: Carlos Ramisch, Aline Villavicencio (carlos.ramischATlif.univ-mrs.fr – avillavicencioATinf.ufrgs.br)

The mwetoolkit is a tool for extracting lists of MWEs from texts (http://mwetoolkit.sourceforge.net). The goal is to extend it so that the output is identical to the input corpus, but with the MWEs marked using a special markup. The expected outcome is a fully functional tool that allows to obtain a corpus with token MWE annotation (as opposed to MWE lists independent from the text, as it is currently the case).

Tasks:

Understand the mwetoolkit and run some toy experiments

Define an output format for the markup, something like <mwe id=”1″>get</mwe> it <mwe id=”1″>off</mwe>

Develop a tool “project.py”, which takes as an input the corpus and the list of MWEs (possibly obtained using candidates.py with -S option) and projects/annotates the MWEs on the corpus.

Evaluate the annotation on a sample corpus of phrasal verbs in English

3) Extending the MWE annotation of parallel corpora to noun compounds/idioms

Supervisors: Carlos Ramisch, Aline Villavicencio (carlos.ramischATlif.univ-mrs.fr – avillavicencioATinf.ufrgs.br)

The project has already some parallel data on phrasal verbs. The idea of this internship would be to annotate the same parallel corpora with other types of MWEs, namely noun compounds and idioms. Noun compounds are sequences of nouns (and/or other elements) that represent a single noun, like telephone booth, wasing machine, cable car. Idioms are expressions which cannot be interpreted semantically word-by-word; we will focus on verb-noun idioms like foot the bill, spill the beans, kick the bucket. The student will use some pre-existing tool like mwetoolkit and the respective patterns. We focus on a small set of expressions. Some manual validation might be required on the test sets created.

4) Modelling the variability of support verb constructions

Supervisors: Carlos Ramisch, Aline Villavicencio (carlos.ramischATlif.univ-mrs.fr – avillavicencioATinf.ufrgs.br)

Support verb constructions like take a picture, pay attention, make a call allow less variability than similar constructions like take a bus, pay money, make a cake. The idea of the internship is to automatically generate variants for support verb constructions and then calculate the entropy of this set of variants. It is based on similar work developed for VPCs (Villavicencio et al. 2008 CONLL). Variability information can be used as a feature for MWE identification and even for better translation.

5) Creating a dataset for phrasa verb compositionality based on wordnet synonyms
Supervisors: Carlos Ramisch, Aline Villavicencio (carlos.ramischATlif.univ-mrs.fr – avillavicencioATinf.ufrgs.br)
Annotating compositionality for a phrasal verb like take off is barely impossible out of context or using a numeric scale or enumeration of classes. However, finding synonyms, paraphrases and equivalents is a more natural task that native speakers and linguists are comfortable with. Our goal is to assign a compositionality judgement to phrasal verbs in context (sentences) implicitly, by asking the annotator to replace it by other verbs.

6) Evaluation protocol for assessing MWEs in Automatic Speech Recognition
Supervisors : L. Besacier & A. Villavicencio (Laurent.BesacierATimag.fr -avillavicencioATinf.ufrgs.br)

This work would consist in building an evaluation protocol for assessing MWEs in Automatic Speech recognition. To the best of our knowledge, it has not been done before. Our idea would be to record a specific read speech corpus in english containing a large number of phrasal verbs. This could be done by having 3 speakers each recording the 500 utterances of LIG-LIDILEM phrasal verb corpus (and recording simultaneously similar sentences without phrasal verbs). After that, LIG ASR system (based on KALDI speech recognition toolkit) would be applied and word error rates (WER) between the PV and non-PV corpora would be compared. If any, differences in performance would be investigated more deeply afterwards.

7) Unsupervised extraction of MWEs from word alignments and integration with MWEtoolkit
Supervisors : C. Ramisch & H. Caseli (carlos.ramischATlif.univ-mrs.fr – helenacaseliATdc.ufscar.br)

The idea is to develop a tool – integrated into the mwetoolkit – that takes as input parallel text automatically word-aligned (e.g. the output produced by GIZA++) and detect patterns that indicate the presence of multiword expressions. The intuition behind this idea is that MWEs are generally not translated word by word. At a first moment, the student can reimplement the simple methods proposed by Caseli et al. (2010) and Tsvetkov and Wintner (2011). Then, he/she can use some contextual measure to detect when some words in source language are systematically aligned with some words in the target language in a specific context and not in other contexts. The focus will probably be on idiomatic constructions since they are generally not translated word by word.

C. PRODUCTION SCIENTIFIQUE CO-SIGNEE AVEC LE PARTENAIRE ETRANGER

a) Liste des publications parues, acceptées ou soumises (préciser) dans des revues avec comité de lecture

b) Liste des publications dans des ouvrages (livres, proceedings, … préciser)

c) Liste des présentations à des colloques co-signées avec le partenaire étranger
(indiquer si exposés oraux ou affiches)

- Bruno Laranjeira, Viviane Moreira, Aline Villavicencio, Carlos Ramisch, Maria José Finatto, "Comparing the Quality of Focused Crawlers and of the Translation Resources Obtained from them", Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, 2014.

- Muntsa Padró, Marco Idiart, Aline Villavicencio, Carlos Ramisch, "Comparing Similarity Measures for Distributional Thesauri", Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, 2014.

- Muntsa Padró, Marco Idiart, Aline Villavicencio, Carlos Ramisch, "Nothing like Good Old Frequency: Studying Context Filters for Distributional Thesauri", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014) - short papers, Doha, Qatar, October, 2014.

d) Liste des brevets en co-propriété

e) Autres co-productions (bases de données, plateformes, sites web, portails thématiques… préciser)

site Web du projet AIM-WEST : http://aim-west.imag.fr

D. OBSERVATIONS EVENTUELLES

La prochaine réunion « physique » du projet est prévue à Porto-Alegre du 21 au 23 Octobre 2015