Université de Lorraine - IDMC
Masters SC & TAL – M1
2020-2021

Fiche de projet tutoré / Project form

Correction de corpus / Corpus correction

Encadrement / Supervisors
Sémagramme, LORIA
Bruno Guillaume (bruno.Guillaume@loria.fr)

Description / Description

The goal of project is to explore the existing tools and methodologies that have been proposed for automatic detection of potential errors in treebanks. Papers in the joined bibliography propose these kind of tools or methods.

The first part of the project will be dedicated to the bibliographic part: make a large inventory of these kind of methods, compare them and decide the ones that are suitable for exploitation on UD corpora.

During the second part, the students will experiments these different methods on a subset of UD corpora (on the languages they know), analyse the output and propose a set of corrections on the corpora. Depending on the availability of an implementation, it will consist in installation and exploitation or it will consist in implementation or adaptation of existing methods.

More generally, the students will have to analyse on which way, methods are similar or complementary and maybe to propose and test their own methods on this task.

Informations diverses : matériel nécessaire, contexte de réalisation /
Various information: material, context of realization

The team Sémagramme participates to the UD project by maintaining mainly French corpora and providing several tools to help corpus improvement.
With this project, the students will participate to these activities.

Livrables et échéancier / Deliverable and schedule

First part: a bibliographic report with a description of the methods and tools with a comparison of the different approaches

Second part: for a selected set of methods and of UD corpora, the analyses of the output of the method and a manual validation on a subset of the identified errors. The final report will conduct a comprehensive comparison of approaches and an evaluation of the benefit obtained with the corpus corrections.

Bibliographie /References (max. 4-5)
[*il ne s'agit pas de la bibliographie complète qui sera fournie aux étudiants au début du projet mais d'une bibliographie indicative pour aider à cerner le sujet*]

Ines Rehbein and Josef Ruppenhofer, "Sprucing up the trees - Error detection in treebanks" CoLing 2018 (https://www.aclweb.org/anthology/C18-1010)

Guillaume Wisniewski, "Errator: a Tool to Help Detect Annotation Errors in the Universal Dependencies Project" LREC 2018 (https://www.aclweb.org/anthology/L18-1711)

Guillaume Wisniewski and François Yvon "How Bad are PoS Tagger in Cross-Corpora Settings? Evaluating Annotation Divergence in the UD Project" NAACL 2019 (https://www.aclweb.org/anthology/N19-1019)

Marie-Catherine de Marneffe, Matias Grioni, Jenna Kanerva, Filip Ginter, "Assessing the annotation consistency of the universal dependencies corpora" Depling 2017 (https://www.aclweb.org/anthology/W17-6514)

Boyd A., Dickinson M., Meurers W. D., « On detecting errors in dependency treebanks », Research on Language & Computation, vol. 6, no 2, p. 113-137, 2008 (https://link.springer.com/content/pdf/10.1007/s11168-008-9051-9.pdf)

Alzetta C., Dell'Orletta F., Montemagni S., Simi M., Venturi G., "Assessing the Impact of
Incremental Error Detection and Correction. A Case Study on the Italian Universal Dependency Treebank", UDW 2018 (http://universaldependencies.org/udw18/PDFs/39_Paper.pdf)

Encadrement / Supervisors

Sémagramme, LORIA
Bruno Guillaume (bruno.Guillaume@loria.fr)

Description / Description

The goal of project is to explore the existing tools and methodologies that have been proposed for automatic detection of potential errors in treebanks. Papers in the joined bibliography propose these kind of tools or methods.

The first part of the project will be dedicated to the bibliographic part: make a large inventory of these kind of methods, compare them and decide the ones that are suitable for exploitation on UD corpora.

During the second part, the students will experiments these different methods on a subset of UD corpora (on the languages they know), analyse the output and propose a set of corrections on the corpora. Depending on the availability of an implementation, it will consist in installation and exploitation or it will consist in implementation or adaptation of existing methods.

More generally, the students will have to analyse on which way, methods are similar or complementary and maybe to propose and test their own methods on this task.

Informations diverses : matériel nécessaire, contexte de réalisation /
Various information: material, context of realization

The team Sémagramme participates to the UD project by maintaining mainly French corpora and providing several tools to help corpus improvement.
With this project, the students will participate to these activities.

Livrables et échéancier / Deliverable and schedule

First part: a bibliographic report with a description of the methods and tools with a comparison of the different approaches

Second part: for a selected set of methods and of UD corpora, the analyses of the output of the method and a manual validation on a subset of the identified errors. The final report will conduct a comprehensive comparison of approaches and an evaluation of the benefit obtained with the corpus corrections.

Bibliographie /References (max. 4-5)
[*il ne s'agit pas de la bibliographie complète qui sera fournie aux étudiants au début du projet mais d'une bibliographie indicative pour aider à cerner le sujet*]

Ines Rehbein and Josef Ruppenhofer, "Sprucing up the trees - Error detection in treebanks" CoLing 2018 (https://www.aclweb.org/anthology/C18-1010)

Guillaume Wisniewski, "Errator: a Tool to Help Detect Annotation Errors in the Universal Dependencies Project" LREC 2018 (https://www.aclweb.org/anthology/L18-1711)

Guillaume Wisniewski and François Yvon "How Bad are PoS Tagger in Cross-Corpora Settings? Evaluating Annotation Divergence in the UD Project" NAACL 2019 (https://www.aclweb.org/anthology/N19-1019)

Marie-Catherine de Marneffe, Matias Grioni, Jenna Kanerva, Filip Ginter, "Assessing the annotation consistency of the universal dependencies corpora" Depling 2017 (https://www.aclweb.org/anthology/W17-6514)

Boyd A., Dickinson M., Meurers W. D., « On detecting errors in dependency treebanks », Research on Language & Computation, vol. 6, no 2, p. 113-137, 2008 (https://link.springer.com/content/pdf/10.1007/s11168-008-9051-9.pdf)

Alzetta C., Dell'Orletta F., Montemagni S., Simi M., Venturi G., "Assessing the Impact of
Incremental Error Detection and Correction. A Case Study on the Italian Universal Dependency Treebank", UDW 2018 (http://universaldependencies.org/udw18/PDFs/39_Paper.pdf)