Demonstrating ConstruKT, a text annotation toolkit for generalized linguistic contructions applied to communication spin

Anna Koroleva[*†], **Patrick Paroubek**[*]

[*] LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
[†] Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands
koroleva@limsi.fr pap@limsi.fr

Abstract

We present ConstruKT, an open and freely available graphical user interface for linguistic annotation of generalized constructions, which are sets of arbitrary constraints on possibly discontinuous multi-word units and relations between multi-word units or other relations. ConstruKT was written in Python with the TkInter graphical library in the context of the MiRoR European project, in order to provide a flexible development NLP workbench for research on automatic communication spin detection in research publication for the domain of the health and bio-medical research. The design aims for ConstruKT were to provide a graphical annotation interface for medical domain experts with no expertise in linguistics or Natural Language Processing (NLP), enabling them to annotate arbitrary relations between overlapping discontinuous mutli-word units in a scientific article, and at the same time to have for the project a versatile light weight NLP development workbench in Python. ConstruKT is easily retargetable to any application domain, since its core functionalities define generic multi-word units and binary relations than can be created on the fly and specialized at will.

1. Introduction

MiRoR[1] is a large international collaborative project devoted to improving the planning, conduct, reporting and peer reviewing of health care research, in the course of which 15 PhDs address various scientific questions related to improving reporting practices. The one for which ConstruKT was created is about automatically identifying communication spin (Boutron et al., 2014), i.e. results beautifying, in scientific reports, letting the reader believe that the results are more important than the experimental observations show. Spin in scientific publications comes

Conclusions: *Modafinil* **may be useful** *in controlling cancer-related fatigue, especially in patients who present with severe fatigue.*

Figure 1: Example of two types of spin, first (in bold) an unjustified positive evaluation of the treatment, and second (underlined) focusing on the results obtained with only a subgroup only of the initial test group.

in different types and is difficult to identify because it involves semantics, pragmatics and argumentation structure. Due to the complexity and heterogeneity of the concept of spin, we are focusing first on publication reporting Randomized Controlled Trials (RCTs) and only on a few types of spin:

1. distorted reporting related to the primary outcome, e.g. not defining the primary outcome in the abstract;
2. distorted reporting related to various types of positive statements regarding the studied treatment , e.g. positive evaluations (including recommendations to use the treatment) despite statistically non-significant results for the primary outcome.

We although limit our first investigations to spin in abstracts, where it is easier to detect and because the impact of our contribution will be maximized. From our study of the literature and interviews of domain experts, we can outline the following categories of supporting evidence for claims in an article: primary and secondary outcomes of the reported study; statistical significance of results for each outcome; population and patients' subgroups assessed; study design; limitations of the study; adverse events observed; various types of positive evaluations of the experimental treatment (stating beneficial effect of the treatment, stating similarity between the experimental and control treatment regarding a positive characteristic).

In order to address the challenge of automatic spin detection, we first defined and annotation scheme for spin in scientific publications (Koroleva and Paroubek, 2018) and built the first annotated corpus with communication spin annotations. From an initial text corpus of 3,938 articles from PMC [2] with the publication type "RCT", we annotated 664 sentences with significance level mentions, the relations between an outcome and a significance level for 2,678 pairs of outcomes and significance level markers (split roughly equally between positive and negative examples) and the text structure (abstract versus body) in all articles. We now present the experiments we did with various annotation software solutions and explain why we built ConstruKT.

2. Existing annotation toolkits

We considered several tools for linguistic annotation of texts : Knowtator (Ogren, 2006), GATE (Cunningham et al., 2013), Brat (Stenetorp et al., 2012), Glozz (Widlöcher and Mathet, 2012) and WebAnnotator (Tannier, 2012). First, we selected Glozz as the best suited for our needs: it is powerful and flexible in regard to types

[1]MiRoR is funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207. See http://miror-ejd.eu

[2]PMC (PubMed Central) is a database of full-text articles in the domains of biomedicine and life sciences. Official site: https://www.ncbi.nlm.nih.gov/pmc/

of units and relations that can be annotated, allowing to annotate all the information relevant for spin with an an ergonomic interface. But after developing a set of annotation guidelines and running a few tests with domain experts, we realized that it did not met our practical requirements, because Glozz provides too rich information which overloads people novice in text annotation task. Then we changed our approach an decided to split the annotation task into a set of simple annotation questions, each associated to a text excerpt, presented to the annotator by means of the LimeSurvey application[3]. But this solution was abandoned, for two main reasons : first, if the use of a web service application solved the problem of installation of the annotation software, nevertheless the overload of work required to format the question into the form a questionnaire was imposing a much too heavy cost penalty for considering annotating a large corpus, and second the expert annotators did not validate splitting of the annotation task into a multiple of disconnected subtasks because they needed to have the whole article available for browsing when annotating text, displayed as much as possible in the same rendering as in the original document. Finally, the solution that fulfilled all the requirements was developing an annotation interface in Python with the TkInter[4] Python interface to Tcl/Tk.

3. ConstruKT Annotation Toolkit

ConstruKT annotation toolkit[5] provides functionalities for annotating text with possibly overlapping and discontinuous multi-word units (MWUs) and binary relations between multi-words units or other relations. Both MWUs and relations can be dynamically typed and decorated with arbitrary feature structures dynamically specified, offering thus versatility during the development of annotation and easy of porting to existing annotation formats. The display of information was deliberately kept minimal in order to keep the display uncluttered and to limit the mental workload of the annotator. The later has only to perform selection of text spans and information slot filling to create new annotations. The MWUs and relations are represented uniquely by means of various text span highlight coloring or change of font color. When annotation overlap, only two overlapping annotations can be displayed at a time, but the number of overlapping annotation is limited only by the memory available. Relations can be grouped into structures called *constructions* in order to easily combine syntactic and semantic constraints, for instance through interfaces to a parser and a semantic network. Annotations are represented and saved as Python source code which spares the developer the writing of import/export procedures and facilitates interfacing with the Python package ecology. For instance the current version uses different NLP packages like: NLTK, TreeTagger, Spacy or Bert.

[3]https://www.limesurvey.org

[4]https://wiki.python.org/moin/TkInter

[5]Download ConstruKT at: https://mycore.core-cloud.net/index.php/s/FuhSBnk2oEJbEji

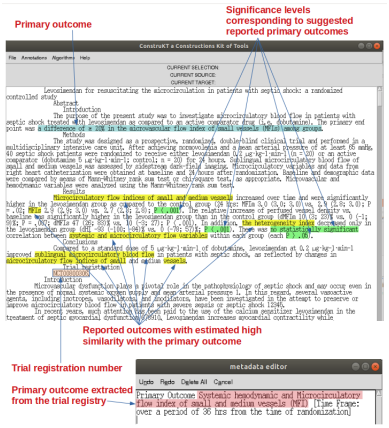

Figure 2: Example of a processed text

4. Acknowledgements

We thank our advisors in the domain of medical reporting, prof. Isabelle Boutron from the University Paris Descartes, prof. Patrick Bossuyt from the University of Amsterdam, and Liz Wager from SideView, for their highly appreciated help in the conduct of this work.

5. References

Boutron, Isabelle, Douglas Altman, Sally Hopewell, Francisco Vera-Badillo, Ian Tannock, and Philippe Ravaud, 2014. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the spiin randomized controlled trial. *Journal of Clinical Oncology*.

Cunningham, Hamish, Valentin Tablan, Angus Roberts, and Kalina Bontcheva, 2013. Getting more out of biomedical documents with gate's full lifecycle open source text analytics. In *PLoS Computational Biology*.

Koroleva, Anna and Patrick Paroubek, 2018. Annotating Spin in Biomedical Scientific Publications : the case of Random Controlled Trials (RCTs). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan.

Ogren, Philip V., 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Morristown, NJ, USA: ACL.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii, 2012. brat: a web-based tool for nlp-assisted text annotation. In *Proc. of the Demonstrations at the 13th EACL*. Avignon, France: ACL.

Tannier, Xavier, 2012. Webannotator, an annotation tool for web pages. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey.

Widlöcher, Antoine and Yann Mathet, 2012. The glozz platform: a corpus annotation and mining tool. In *ACM Symposium on Document Engineering*.