Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

SIDRES : A Novel Annotation Tool For The Automatic Detection Of Semantic Entities

Julieta Murata[1] Rémy Carrette[1, 2] Pierre Jourlin[2]
(1) SATT Sud-Est, Le Silo, 35 Quai du Lazaret, CS 70545 13304 Marseille Cedex 02, France
(2) Laboratoire d'Informatique - Avignon Université, 339 Chemin des Meinajaries, 84000 Avignon, France
julietamurata@gmail.com, remy.carrette@alumni.univ-avignon.fr, pierre.jourlin@univ-avignon.fr

Résumé

Nous présentons un nouvel outil d'annotation nommé SIDRES (Système Interactif de Détection et de Reconnaissance d'Entités Sémantiques). SIDRES fournit un environnement d'annotation pour la classification d'unités textuelles à partir de catégories *ad hoc*. Ces catégories peuvent être associées à des contextes comme un moyen de désambiguïsation d'unités identiques appartenant à des catégories différentes. SIDRES a été développé dans le cadre d'un partenariat industriel visant à effectuer un transfert de technologie du laboratoire de recherche publique vers des acteurs de l'industrie.

Abstract

We present a novel annotation tool called SIDRES (Système Interactif de Détection et de Reconnaissance d'Entités Sémantiques [Interactive System for the Detection and Identification of Semantic Entities]). SIDRES provides an annotation environment for classifying text units through *ad hoc* categories. These categories can be coupled with contexts, so as to provide a means for the disambiguation of formally identical units assigned to different categories. SIDRES was developed as part of an industrial partnership between the LIA (Laboratoire d'Informatique d'Avignon [Research Institute of Informatics at the University of Avignon]) and a French company in the eHealth sector. This partnership was created within the framework of a technology-transfer project promoted by the SATT Sud-Est, whose core mission is bringing together industry and research institutions.

Mots-clés : rapports cliniques, annotation, grammaire ambigüe locale
Keywords: medical reports, annotation, locally ambiguous grammar

1 Theoretical Aspects

SIDRES relies on a particular text data structure called "confusion tree". Confusion trees allow to define locally ambiguous grammars capable of representing multi-word expressions, determine their boundaries, and provide means for disambiguation as well as group these expressions according to different syntactic or semantic categories based on their linguistic context. This type of structure is exploited by an algorithm in the line of Tomita's Generalized LR (Tomita, 1984). The algorithm allows to extract and disambiguate concurrent terminology in natural language texts.

The example below shows the term "Paris" being categorized as both a location (the capital city of France) and a person (Paris Hilton). Subsequent linguistic contexts allow for the disambiguation of these terms.

```
Paris (Location : 111 ; Person : 15)
|____ de Paris (Location : 47, Person : 5)
|        |____ville de Paris (Location : 47)
|____ Paris lance (Location : 20, Person : 3)
|        |____ Paris lance un (Location : 10, Person  : 1)
|                 |____ Paris lance un audit (Location : 5)
|____ Paris, (Location : 100 ; Person  : 12)
         |____ Paris, considérée (Location : 80 ; Person  : 5)
                  |____ Paris, considérée comme (Location : 79 ; Person : 5)
                           |____ Paris, considérée comme une (Location : 75 ; Person : 4)
                                    |____ Paris, considérée comme une jet-setteuse (Person : 2)
```

2 Functionalities

2.1 Interface Design, Categories, and Contexts

SIDRES is coded in Python3 and uses the GTK4 framework. The interface design has a two-panel division: the left-hand panel provides for the creation of annotation categories, and the right-hand panel consists of a text visualization window. In order for the annotation corpus to be displayed on the right, the corpus must first be loaded either by connecting to a local database or by selecting a text file from the local file system. Annotators can then manually define and edit a set of categories of their choice (e.g.: <person>, <location>) according to their own annotation model, and assign a distinct color to each category. Upon manually selecting and right-clicking on textual units, the user can simply add the unit to the intended category. If a unit is annotated under more than one category (e.g.: the term “Paris” is a <person> and a <location>), users can select a linguistic context so as to provide a means of disambiguation. At any point in the annotation process, users can save and export their work in a JSON file, and load it up again.

2.2 Other Features

To satisfy the goals of annotation management, the upper menu allows for further functions, namely: (1) a display function that presents the list of categories and a cumulative list of annotated units; (2) a research function to find identical units that have been assigned to different categories; (3) basic statistics on the annotated units (total number of units, categories, distribution graphs).

2.3 Use Examples in the Medical Domain: the Negation of Signs and Symptoms, and the Negation of Medical History in Medical Reports in French

We present two case examples of corpus annotation of medical reports in French using SIDRES.

In medical records, information organization follows a stable pattern and generally conforms to the SOAP model (such as age, sex, signs and symptoms, medical history, test results, etc). Introducing polarity in such typologies can lead to more fine-grained distinctions that have great informational value for clinical and research purposes. In fact, negation is a major source of poor precision in medical information retrieval systems (Rokach, Romano & Maimon, 2008; Averbuch, Karson, Ben-Ami, Maimon & Rokach, 2004).

Once we display a corpus of free-text clinical reports on SIDRES, we can create a set of twin categories : [1] <signs_symptoms> and [2] <signs_symptoms_n>, for encoding the presence and the absence of signs and symptoms, respectively ; [3] <history> and [4] <history_n>, for encoding the presence and the absence of medical history. We then incorporate the terms in the MeSH “C” descriptor to populate the SIDRES categories. While all four lists now feature the exact same content, introducing contexts on SIDRES can help solve this contradiction. We apply the French version of the NegEx patterns (Chapman, Bridewell, Hanbury, Cooper & Buchanan, 2001) for [2] (i.e., sans/pas de/absence de <signs of symptoms>). For [3] and [4], we introduce corpus-based syntactic patterns such as “with a history of" and “with no history of”. [1] is left as the context-free, unmarked category.

3 Conclusion

We presented SIDRES, a new annotation tool that allows for the creation of categories and the introduction of contexts. We demonstrated the value of contexts for addressing negation in medical reports in French. Thanks to its flexibility, simplicity of use, and broad application possibilities, SIDRES can be easily incorporated to varied, small and large-scale annotation projects. SIDRES can be used for research purposes under a non-commercial license with an agreement of Avignon University. For commercial purposes, please contact Guillaume Gouvernet (guillaume.gouvernet@sattse.com).

References

AVERBUCH, M., KARSON, T., BEN-AMI, B., MAIMON, O. & ROKACH, L. (2004).Context-sensitive medical information retrieval. In Proc. of the 11th World Congress on Medical Informatics (MEDINFO-2004), pages 1–8. Citeseer. DOI : 10.3233/978-1-60750-949-3-282

CHAPMAN, W., BRIDEWELL, W., HANBURY, P., COOPER, G., BUCHANAN, B . (2001). A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries, Journal of Biomedical Informatics, Volume 34, Issue 5, pages 301-310, DOI: 10.1006/jbin.2001.1029.

ROKACH, L., ROMANO, R. & MAIMON, O (2008). Negation recognition in medical narrative reports. Inf Retrieval 11, 499–538. DOI : 10.1007/s10791-008-9061-0

TOMITA M. (1984). LR parsers for natural langages. *10th International Conference on Computational Linguistics*. *COLING*: 354-357. DOI : 10.3115/980491.980564