

A Linguistic Annotation Facility for Pragmatic Features in an Authoring Tool

Jakub Gawryjolek
jgawry@gmail.com

School of Computer Science
University of Waterloo

Supervisor: Chrysanne DiMarco

Linguistic annotation provides additional information asserted with a particular purpose in a document or other piece of information. It is widely used in various fields, from computing and bioinformatics, through imaging, to law and linguistics. The purpose of this abstract is to present our annotation tool, which is capable of annotating HTML documents with a configurable list of pragmatic features. To this point it has been used for annotating the rhetorical purpose of scholarly citations, but it can be easily extended to annotate textual data with any type of pragmatic feature. The annotation facility is a component of our Web-based document authoring tool, developed within our IN³SCAPE project at the University of Waterloo.

Citations in scholarly articles play an important role in creating relationships among mutually relevant articles within a research field by expressing semantic links between the documents. These inter-article semantic relationships represent the argumentation structure intrinsic to all scientific writing. Therefore, determining the nature of the exact relationship between a citing and cited paper requires an understanding of the rhetorical relations within the argumentation context in which a citation is placed. To determine these relations automatically in scientific articles, we have proposed that associated pragmatic features within the context of a citation may be automatically determined by computational linguistic analysis. In our project, our goal is to automatically annotate the purpose of a citation, on the basis of these pragmatic features, using a combination of discourse analysis and machine learning techniques.

Our project is concentrating on several important research problems in which the annotation tool can be usefully applied. Linguistic annotations can be used to indicate general semantic relationships between a source document and a target document. For example, within the scope of the IN³SCAPE project, citation annotations provide crucial information about semantic relationships between two scholarly articles, e.g., whether the knowledge claims in the document being cited support, contradict, mention, etc., the claims or other ideas in the citing paper. Provided that all the documents in the collection have been annotated, we could create a database in the form of a “typed citation index” of the annotation documents. Such a database might be queried and from a researcher’s point of view the result of a query on such a database might be very valuable and would improve searching possibilities for related materials. Another problem that our annotation tool will be extended to handle is the automated annotation of classical rhetorical figures. Rhetorical figures are patterns of text in which a characteristic form modifies the standard meanings of words and leads to a change or an extension of meaning. Assigning the attributes to these figures and collecting a substantial number of annotated documents can provide crucial information for the understanding of whole documents and automated rhetorical figure classification.

A significant implementation feature of the annotation tool is that all information concerning annotations are saved in a separate XML file. Therefore, the original text is never modified. Moreover, the tool allows modification of pre-existing annotations for certain documents so that the descriptions can be edited at any time. Lastly, each manual annotation records its human annotator, together with a certainty value indicating how sure the person is of the type of the annotation.

The general purpose of the created tool is to annotate all types of linguistic annotations. Moreover, it is usable not only for annotating but also for authoring. Navigation through the existing annotations and the creation of new ones in the tool is simple. The former can be done in either of two ways: the user can either select the existing annotation by pointing at the place in the HTML text or she can use the navigation

buttons. When the desired text is selected, the user can then indicate which pragmatic features apply to the selected text. The user can also provide a description of the pragmatic cues, which serves as an additional comment in the annotation. All updates are recorded instantly. Adding new annotations is performed by simply marking a piece of text. In order to provide these features some implementation issues had to be resolved. For example, the HTML documents displayed in our tool differed from the original ones (some edition-related characters have been added by Java components). Therefore, we had to implement an alignment tool between these documents. Secondly, we had to provide a means of saving and re-reading the annotations, which included the creation of XML documents with cross-references (REFIDs), parsing them and finally displaying in a human-readable form.

As mentioned, our annotation tool can be used in the future for the automated classification of general semantic links between documents. In our IN³SCAPE project it is being used for automated citation annotation but our overall goal is to develop an annotation tool that will create mappings between concepts represented in an annotation schema and any type of citation. Ideally, it will be adaptable to annotating general semantic links between online documents. This feature will be extremely useful for the developing significance of the Semantic Web.