

CLAU – A Service-Oriented System for Complex Language Alignment: Architectural Aspects

Claudiu Mihăilă, Corina Forăscu, and Sabin C. Buraga

Abstract—In the last years, parallel corpora have become an effective framework to study how well the linguistic phenomena and, more specifically, annotation schemata can be applied when importing the annotations from one language to the other(s). In the case of automatic import, the evaluation and correction are better to be performed by linguists using specific software. The paper proposes CLAU – a service-oriented interactive application allowing users to import, evaluate, correct, and share XML-based annotations in parallel texts. The design, general architecture, and implementation are discussed. Also, two use cases are presented: temporal annotations in parallel texts and how CLAU facilitates social Web interactions between language scientists.

Index Terms—Parallel text processing, cross-language studies, service-oriented architecture.

I. INTRODUCTION

DUe to their extensive use in many NLP applications, in comparative language study, and their importance in language industries, parallel corpora are continuously created, improved and exploited [20]. Linguistic resources in a target language can be easier developed based on the linguistic knowledge (annotations) encoded in a source text, and using multilingual technologies – such as alignment at various text levels and annotation transfer –, provided that the appropriate text preprocessing tools exist for the source language. In a multiple language setting, it is easier and cheaper to correct automatically imported annotations than to create them from scratch. The individual language users should then use either a set of language specific rules, or directly to manually correct the annotations imported from the source language. Recent experiments with transferring annotations – word senses, syntactic and semantic relations, collocations [18], temporal information [9], coreference chains [15] – show that word alignment technology is a successful solution for the transfer of information from the tokens in one language to their translation equivalents in the other language.

Manuscript received October 21, 2008. Manuscript accepted for publication February 19, 2009.

Claudiu Mihăilă is with Faculty of Computer Science, “A.I.I. Cuza” University of Iași, 16, General Berthelot, 700483, Iași, Romania (e-mail: claudiu.mihaila@info.uaic.ro).

Corina Forăscu is with Faculty of Computer Science, “A.I.I. Cuza” University of Iași, 16, General Berthelot, 700483, Iași, Romania and with the Romanian Academy Research Institute for Artificial Intelligence, Romania (e-mail: corinfor@info.uaic.ro).

Sabin C. Buraga is with Faculty of Computer Science, “A.I.I. Cuza” University of Iași, 16, General Berthelot, 700483, Iași, Romania (e-mail: busaco@info.uaic.ro).

Based on the existing NLP tools, briefly presented in Section II, *CLAU – the Complex Language-Alignment User-oriented system* – will be presented, available as an open-source collection of Web services working with XML-annotated parallel corpora. The users can configure CLAU according to their needs when developing and/or correcting parallel annotations, and they can collaborate in a cross-lingual and cross-cultural environment.

The paper describes – in Sections III and IV – the design, the general architecture, and several implementation solutions of the proposed system. CLAU is developed as an interactive application allowing users to import, correct, and evaluate annotations in parallel texts. The CLAU parallel text alignment system will use certain techniques implemented as local or external Web services, following the Service-Oriented Architecture (SOA) methodology described in Section 3 of this paper.

Section V presents two case studies proving how well the annotator can handle and/or improve annotations in parallel texts when using CLAU. These case studies show the impact of using a system like CLAU in the activities related to the annotations of temporal information in parallel texts, and in the support for social Web interactions. By using CLAU, less-specialized users of NLP tools are able to surmount different difficulties and redundant work regarding parallel corpora.

The paper ends by presenting the main conclusions and also gives important future research directions.

II. STATE OF THE ART

Next to the abundance of linguistic resources, there are many specific tools for their exploitation and use [20]. Depending on the type of resource to work with – be it monolingual, parallel or comparable corpora, lexicons or dictionaries of various types, the existing tools have different characteristics:

- Can be used locally or remotely (through a server),
- Can accept various file-formats or annotation standards,
- Can permit or not a collaborative and/or cross-lingual work,
- Can or cannot be configured according to the user’s needs,
- Can have different types of licenses, which have impact on their use and further development by other members of the research community.

