A GATE Plugin For Tagging French Medical Texts with UMLS concepts

Thierry Delbecque, MSc, Pierre Zweigenbaum, PhD
LIMSI, CNRS, F-91403 Orsay, France

Abstract

GATE is an open source JAVA platform for annotating texts. Its plug-in architecture makes it easy to add new components, and to organize computations by assembling these components into work-flows. In an effort to provide non English languages with tools that are necessary in modern medical information processing, we present a specialized GATE component (MetaCoDe V0.2) that aims at tagging medical corpora available in some non English languages (French being the first targeted one) with UMLS concepts.

Motivation

Popular tools exist nowadays for extracting UMLS concepts from medical texts. The more famous of these tools may be MetaMap,1 another one being MGrep, from Michigan University. Besides, GATE (General Architecture for Text Engineering, http://gate.ac.uk) is a rich widely used JAVA platform that allows to leverage NLP algorithms in order to create complex text processing applications. A wrapper has been recently developed around the MetaMap JAVA API in order to make it usable as a GATE plug-in, thus still spreading its potential usage (http://www.orbitproject.org/resource/metamap-plugin-gate-framework). Unfortunately, English is the only language that could benefit from such initiatives: it has been recognized that MetaMap can be applied to English texts only because its implementation and syntactic algorithms are especially tailored for this language.1 Other languages such as French or Spanish suffer from the lack of similar tools.

MetaCoDe V0.1 (http://www.semantic-valley.org/en/metacode/guide.pdf) is an early attempt to remedy this situation. It is made of a set of PERL and C++ open source programs that use little natural language processing to find occurrences of UMLS concepts in English or French texts. An evaluation we made of the English setting of MetaCoDe V0.1 using a corpus of 9,200 words and the output of MetaMap as the gold standard estimated a precision of 0.93 and a recall of 0.76.2 Unfortunately this tool is not easy to use, as it demands a fair amount of preprocessing to be done to run the tagger. As another drawback it cannot be seamlessly integrated into wider processing chains.

To overcome these difficulties, we are developing MetaCoDe V0.2 (http://sourceforge.net/projects/metacode) as a new GATE component. The core of MetaCoDe V0.2 is the C++ concept extraction algorithm inherited from V0.1, but it benefits from GATE’s NLP components. Along with tokenization and part-of-speech tagging, noun phrase chunking is achieved thanks to a transducer defined with a JAPE grammar. This makes extending MetaCoDe to a number of languages other than French possible, as it should be “only” a matter of changing the POS tagging rules and rewriting the JAPE grammar. The work-flow should remain the same and the core algorithm of the tagger which is based on lattice exploration does not depend of the application language. A visualization plug-in is being developed in order to browse the output of MetaCoDe V0.2. It allows for instance to interactively query an UMLS database while exploring the tagged corpus.

Planned works aim at improving performances, leveraging the tagging for more specialized information extraction tasks and for subsequent text mining usages, and at serving more languages, Spanish at first.

References
