THE LINGUASTREAM PLATFORM

Frédérik Bilhaut
GREYC, University of Caen, France
fbilhaut@info.unicaen.fr

Resumen: Este artículo presenta la plataforma *LinguaStream*, desarrollada para ofrecer un entorno de desarrollo que permite concebir sistemas complejos de Procesamiento del Lenguaje Natural (PLN). Utiliza extensivamente XML, y demuestra un uso práctico de este estándar y de otros estandars que se apoyan en él para el PLN, insistiendo particularmente en preocupaciones semánticas.
Palabras clave: Procesamiento del Lenguaje Natural, Entorno de Desarrollo, Semánticas

Abstract: This paper presents the *LinguaStream* platform, developped to offer an integrated development environment for designing complex natural language processing (NLP) systems. It relies extensively on XML, and demonstrates a practical use of this standard and surrounding ones for NLP, taking particular attention on semantic concerns.
Keywords: Natural Language Processing, Integrated Platform, Semantics

The LinguaStream platform offers an integrated development environment for designing Natural Language Processing (NLP) systems, especially targeted to semantics-oriented concerns. It relies on the paradigm of iterative enrichment of electronic documents, and provides a comfortable yet powerful way to design complex processing streams, where each step may produce new annotations to be integrated in the document, that may in turn be used by further steps of the processing stream (or by another stream).

Since the platform makes no assumption about the nature of each processing component, the resulting processing streams can be very heterogeneous. This is often necessary in practice, since NLP systems usually involve varied but strongly interdependant processing steps, including : lexical, syntactic, semantic, or discourse-level analyses, document engineering techniques (document importation, transformation or visualisation), or structured data management (involving both storage and retrieval aspects, for example on lexical or semantic databases). The platform addresses this issue, and can therefore be used to produce arbitrarily complex streams, leading to fully functional NLP applications. All produced annotations are represented as feature structures, that are automatically integrated in analysed documents, and made available to subsequent steps of the processing stream.

The platform is composed of an extensible set of NLP-oriented components accessible through

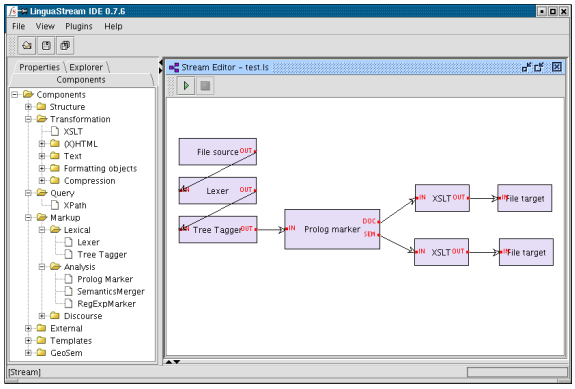

Figure 1: A part of LinguaStream's GUI

a Java™API, as well as a Graphical User Interface (GUI) shown on Fig. 1, allowing processing streams to be visually designed and tested. Although the platform is mainly targeted to computer scientists working on computational linguistics, the GUI makes it also valuable for individuals having minimal technical knowledge in this domain. It can for example be used by linguists to experiment hypotheses on corpus. The main part of the GUI is the stream editor, were processing components can be visually assembled, but the GUI features many other tools, in order to form an integrated environment where all necessary steps of the building of a complete NLP system can be achieved.

LinguaStream makes extensive use of the eX-

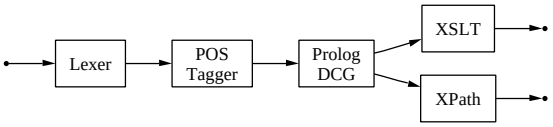

Figure 2: A Processing Stream Sample

"From may 1985 to september 1994"

$$\left[\textit{period} : \left[\begin{array}{l} \textit{start} : \left[\begin{array}{l} \textit{month} : 5 \\ \textit{year} : 1985 \end{array} \right] \\ \textit{end} : \left[\begin{array}{l} \textit{month} : 9 \\ \textit{year} : 1994 \end{array} \right] \end{array} \right] \right]$$

Figure 3: Temporal phrase with its abstract semantics represented as a feature structure

tended Markup Language (XML) as well as other standards that surround it (including RDF, XSchema and XSL). These standards are used at several levels to ensure interoperability beween components, structured data management, and result visualisation. The core of the platform is implemented in Java™, but also makes use of other technologies. For instance, integration with Prolog offers a comfortable environment for designing unification grammars (Covington, 1994), and external tools can be used to perform specific tasks, like TreeTagger (Schmid, 1994) for part-of-speech (POS) tagging and lemmatisation.

Figure 2 shows a simple example of a NLP processing stream that could be built with LinguaStream. For instance, this stream could be used to build a temporal phrases analyser that would mark them in the processed document and generate semantic representations for each of them, like examplified in Fig. 3. First, the input document is processed by a tokenizer to separate words, that are then tagged by a POS tagger. Then, a local unification grammar implemented by a Prolog definite clause grammar (DCG) proceeds to both syntactic and semantic analyses of temporal phrases. Finally, an XSLT stylesheet is used to render the resulting document for visualisation in HyperText Markup Language (HTML), where temporal phrases are highlighted and linked to the visual representation of their semantics. The stream also applies an XPath query on semantic values, in order to filter them according to a given criterion – for example, it could filter temporal phrases referencing periods before a given date.

The LinguaStream platform is mainly developped for the needs of the GeoSem project[1], which unites researchers in linguistics, geography and computational linguistics around the question of semantic analysis of geographical documents. The platform has been successfully used to design and test composite NLP systems described in (Bilhaut, 2003). The resulting process includes syntactic and semantic analysis of spatio-temporal expression, terminology-related and semantic knownledge management systems, statistical methods such as described in (Ferret, 1997), as well as discourse-level analysis mostly based on Charolles' discourse universes model (Charolles, 1997). The platform was also used to develop a web-based search engine based on this technology, allowing efficient information retrieval in geographical information databases.

The platform is still under a constant development process, and we especially hope that collaboration between linguists and computer scientists will lead to the achievement of an easy-to-use yet powerful application. The LinguaStream software is freely available for individual, non-commercial use. More information and downloads are available from the web at `http://www.info.unicaen.fr/~fbilhaut`.

Bibliografía

Frédérik Bilhaut, Thierry Charnois, Patrice Enjalbert, Yann Mathet, 2003, Passage extraction in geographical documents, New Trends in Intelligent Information Processing and Web Mining, Zakopane, Poland.

Michel Charolles, 1997, L'encadrement du discours - Univers, champs, domaines et espace, Cahier de recherche linguistique, 6.

Michael A. Covington, 1994, GULP 3.1: An Extension of Prolog for Unification-Based Grammar, Research Report, AI.

Olivier Ferret, Brigitte Grau, Nicolas Masson, 1997, Utilisation d'un réseau de cooccurrences lexicales pour améliorer une analyse thématique fondée sur la distribution des mots, Actes 1ères Journées du Chapitre Français de l'ISKO, Lille, France.

Helmut Schmid, 1994, Probabilistic Part-of-Speech Tagging Using Decision Trees, International Conference on New Methods in Language Processing. Manchester, UK.

[1]Collaboration between GREYC, ESO (Caen), ERSS (Toulouse), EPFL (Lausanne), supported by the CNRS program "Société de l'Information".