Resumen del proyecto LSGRAM

Toni Badia (tbadia@upf.es), Marta Carulla (carulla@upf.es), Maite Melero (maite@gilcub.es)

1. Antecedentes

En los últimos años la Comunidad Europea está auspiciando una plataforma de elaboración de gramáticas, llamada Alep. Con su desarrollo, así como con el de los proyectos directamente lingüísticos financiados también por ella, la CE pretende relanzar el frente europeo en el área de la lingüística computacional después del relativo fracaso del proyecto Eurotra de traducción automática (1978-92).

Unos de los frentes en los cuales está invirtiendo la CE es el de la creación de recursos gramaticales en el nuevo formalismo. Se trata de crear gramáticas a partir de unos criterios de generalidad y adecuación que las hagan aptas para ser utilizadas posteriormente (enteramente o en parte) en las aplicaciones que se considere conveniente. Ello es particularmente posible porque el formalismo Alep nace deliberadamente como perteneciendo a los corrientes centrales actuales de la lingüística computacional y con la intención de ser básicamente simple y eficiente.

Uno de los primeros pasos realizados en esta direccción fue el dado por el proyecto ET10/52 sobre la reutilización de los recursos gramaticales. En él se estudió la validez de migrar gramáticas de un formalismo a otro (Arnold et al., 1992). Al mismo tiempo se establecieron los primeros pasos en el desarrollo de gramáticas para el alemán, el inglés y el español (Badia,1993). A juicio tanto de los miembros investigadores en el proyecto, como de los responsables de la CE, los resultados de esta primera implementación fueron claramente insuficientes debido a la falta de recursos en el proyecto. Por otro parte, desde la finalización del proyecto ET10/52 la plataforma Alep ha sufrido cambios importantes, de manera que las posibilidades expresivas han aumentado considerablemente.

Todo ello hizo conveniente la puesta en marcha de un proyecto dedicado directamente a la redacción de gramáticas relativamente grandes en las tres lenguas mencionadas antes. Este proyecto es el conocido por el acrónimo LSGRAM, cuyo nombre completo es "Large-Scale Grammars for EC

Languages". Se trata de un proyecto financiado enteramente por la Comunidad Europea bajo el programa LRE ("Language, research and engineering") (su identificador oficial es LRE-61029). Como todos los de esta naturaleza, es un proyecto colectivo, en el cual participan los siguientes centros de investigación:

Institut für Angewandte Informationsforschung (IAI), Saarbrücken, Alemania
University of Essex, Colchester, Reino Unido
Universitat de Barcelona, Barcelona
Universitat Pompeu Fabra, Barcelona
Universität Stuttgart, Alemania

En concreto el proyecto pretende la creación de recursos gramaticales de alto nivel, que puedan ser usados en aplicaciones reales. Por ello, constituye también un banco de pruebas para el formalismo de Alep y su entorno.

El punto de partida de proyecto está en la doble consideración siguiente. Por una parte, los sistemas de procesamiento del lenguaje natural disponibles en el mercado actualmente están faltos de propiedades de alto nivel (como modularidad, estabilidad, extensibilidad, estandarización o facilidad de mantenimiento), que se consideran básicos para conseguir unos niveles mínimos de calidad. Por otra, los centros de investigación han desarrollado sistemas experimentales que parecen satisfacer estos requerimientos; se trata de sistemas basados en los planteamientos modernos, como las gramáticas de unificación. No obstante, estos sistemas (a pesar de llevar muchos de ellos más de diez años de vida) no han servido para crear bancos de datos gramaticales de nivel medio o grande, que puedan ser utilizados posteriormente en aplicaciones diversas.

2. Desarrollo del proyecto

El proyecto LSGRAM empezó a desarrollarse en enero del 1994 y está previsto que finalice en diciembre del 1995. Conceptualmente pueden dividirse en cuatro fases con los contenidos siguientes:

a. fase de preparación dedicada a la definición de la representación de análisis final, investigación de las posibilidades de reutilizar recursos gramaticales ya existentes, selección de un corpus adecuado y construcción de un inventario de los fenómenos que aparecen en él.

b. fase de implementación de gramáticas de cobertura básica para las tres lenguas mencionadas y experimentos de reutilización.

c. fase de ampliación y optimización de las gramáticas básicas, basada en los resultados del estudio de corpus. La cobertura de las gramáticas está orientada al tratamiento de textos 'reales' y por ello la implementación debe incluir detalles incomodos, como las fechas, cifras, abreviaturas, acrónimos, etc.

d. fase de documentación y demostración: redacción de la documentación que acompañan las gramáticas y en especial de un 'manual de codificación de reglas', que especifica algunos de los principios básicos para el diseño de sistemas de tipos y rasgos, el uso de macros, etc. con amplia ejemplificación.

3. Estado actual del proyecto

Hasta el momento se ha realizado la fase preparatoria y está a punto de finalizar la primera fase de implementación de las gramáticas. Los objetivos alcanzados pueden resumirse de la siguiente manera:

1. la representación final del análisis sigue la propuesta de la 'semántica de situaciones' (situation-semantics) de Pollard & Sag (1992) con algunos cambios y extensiones.

2. la reutilización de gramáticas ya existentes es posible parcialmente, las partes más reutilizables de las gramáticas de ALVEY tools para el inglés son la teoría de rasgos, la información de subcategorización , la gramática a nivel de palabra y las reglas para el análisis ortográfico.
Para el alemán se tomaron las gramáticas de CAT2, cuya información léxica y en especial el diccionario de morfemas han resultado ser buenos candidatos para la migración.
la reutilización de las gramáticas ET para el español ha resultado ser difícil debido a las grandes diferencias de expresividad entre el formalismo ET y ALEP.

3. A partir de un estudio previo de corpus, basado en artículos periodísticos sobre temas económicos, se elaborá una lista de fenómenos gramaticales ordenados por su prioridad, determinada a su vez por la frecuencia de aparición del fenómeno en dicho corpus.

La implementación de la gramática de análisis del español en el formalismo ALEP, ha tenido en cuenta esta lista de prioridades, y la cobertura gramatica -en fase de ampliación hasta alcanzar los objetivos de la fase 3- queda actualmente definida de la siguiente manera: en cuanto a esquemas oracionales básicos, se incluyen las relaciones de concordancia, la transformación de la pasiva, los distintos esquemas de subcategorización del verbo en castellano, el fenómeno del control, la construcción del grupo verbal, simple y compuesto, incluyendo perífrasis aspectuales; el grupo nominal con modificadores adjetivales y preposicionales y determinación simple así como oraciones de relativo.

Dicha gramática está inspirada en la teoría HPSG y por lo tanto está claramente lexicalizada, las entradas léxicas son muy ricas en información mientras que las reglas gramaticales son pocas y muy generales. Consta de un analizador morfológico basado en reglas de dos niveles y de un analizador sintáctico propiamente dicho que consulta dos diccionarios en dos fases sucesivas, uno que contiene información de tipo sintáctico y otro que contiene información semántica.

Informes publicados hasta el momento:

Theofilidis,A., Verhagen,M., Badia,T. (1994): "Specifications for a Common Semantic Representation Format".

Verhagen,M. (1994): "Migrating the Alvey Natural Language Tools".

Schmidt,P. (1994): "Migrating CAT2-Based Resources".

Badia,T., Carulla,M. (1994): "Investigation on reusability of ET-ES grammatical resources".

Melero,M. (1994): "Spanish Corpus Study".

Caroli,F., Maas,D., Schmidt,P., Theofilidis,A., (1994): "Investigation of the German LS-GRAM corpus.

Verhagen,M. (1994): "English Corpus Study"

Estos informes se pueden pedir a Toni Badia (tbadia@upf.es).

Bibliografía

Markantonatou, S., Sadler,L. eds., (1994): "Grammatical Formalisms: Issues in Migration", Studies in Machine Translation and Natural Language Processing, Volum 4, Luxemburgo.

Badia, T. (1993): "Inicios de una gramática para el español en ALEP, un formalismo de unificación", Boletín de la SEPLN, N°14, Santiago de Compostela.

Pollard,C., Sag,I. (1992) Head-Driven Phrase Structure Grammar, Final Draft, June 15, 1992.

TRADE (MLAP93/003)

Nuria Bel Rafecas
Maite Melero Nogués
GILCUB (FBG/Universitat de Barcelona)

Introducción

La Traducción Automática se considera una de las aplicaciones más importantes de la Ingeniería Lingüística, desde el punto de vista comercial.
A pesar de que el problema de la TA está lejos de haber sido resuelto, se pone de manifiesto la necesidad de disponer de productos operativos que cubran, al menos parcialmente, la demanda del mercado en este sentido, proporcionando ayudas y herramientas para la traducción.
Así, desde instancias europeas, se favorece el apoyo a los proydctos aplicados de TA que sean capaces de reutilizar los recursos y las tecnologías existentes, con objeto de crear productos industriales aptos para ser comercializados.
TRADE está basado en un sistema prototipo de traducción automática, conocido como E-STAR, desarrollado bajo los auspicios del programa EUROTRA, y que supone una extensión y mejora del propio prototipo de EUROTRA, capaz de servir como base para un sistema industrial. El proyecto TRADE pone énfasis en los requerimientos principales del usuario, como son, rendimiento, extendibilidad, robustez, asequibilidad y tratamiento de "textos

reales". El sistema ha sido especialmente diseñado para superar los problemas propios de los sistemas estratificacionales (según el modelo EUROTRA, basado en la transferencia entre niveles generadores), pero conservando, sin embargo, sus características más deseables. Así, el formalismo lingüístico, basado en la unificación, siendo totalmente compatible con el de EUROTRA, está dotado de mayor poder expresivo e incorpora mecanismos de "seguridad" en la traducción, lingüísticamente motivados, que permiten alcanzar la deseada robustez (relajación de condiciones, reglas de preferencia, ...).

Descripción del proyecto

El proyecto TRADE, se centra, en la actualidad, en tres lenguas europeas y en la traducción de cualquiera de ellas a las otras dos. Estas son: inglés, español e italiano. Se prevee que el sistema sea capaz de traducir textos pertenecientes a ámbitos limitados, definidos por los propios usuarios finales. Las áreas sobre las que actualmente se está trabajando son: la Seguridad Social y la informática de usuario.

Participan en el proyecto los siguientes centros de investigación:

Centro de Cálculo de Sabadell, Barcelona (España)

Gruppo DIMA, Turín (Italia)

Universidad de Manchester, Manchester (Reino Unido)

GILCUB (FBG/Universitat de Barcelona), Barcelona (España)

El proyecto se inició en enero de 1994 y su desarrollo está organizado de la siguiente manera:

Fase 1: Definición de especificaciones sobre la arquitectura general del sistema y sobre la cobertura lingüística, selección de corpus y elaboración de baterías de prueba.

Fase 2: Desarrollo del software básico y de las herramientas de procesamiento de texto. Migración de los recursos lingüísticos de las gramáticas de Eurotra al nuevo formalismo. Optimización de las implementaciones en base a las características innovadoras del E-STAR. Cobertura extensiva de los fenómenos según el estudio de corpus de la fase previa. Desarrollo de la interfície de usuario. Integración de todos los módulos.

Fase 3: Verificación, validación y documentación del sistema.

Hasta este momento se ha cubierto la primera fase y, prácticamente, la segunda, habiéndose alcanzado, en líneas generales, los objetivos propuestos. El proyecto fue sometido recientemente a la evaluación externa por parte de expertos en el ámbito de la TA, con resultados altamente satisfactorios.

Informes publicados

Espejo J.M., Pérez F., Somá E. "System architecture and design specifications".

Mazzini G., Steinberger R., Melero M. "Corpus study and coverage definition".
Allegranza V. "Criteria for testing and evaluation of linguistic components".
Espejo J.M, Pérez F., Chambers C. "Text Handler".
Somà E., Tesio R. "Core Software Enhancement Specification".