El Corrector de Planeta

Yolanda Castellón, Albert Collet, Ana Gonzalvo, Xavier Lloré,
Ágata Ortega, Gema Pérez, Jordi Pérez, David Trotzig
Planeta Actimedia, S.A.
Banco de Contenidos
Departamento de Lingüística Computacional
mailto:dtrotzig@planeta-actimedia.es

Resumen Aquí se muestra parte de la tecnología que se está desarrollando en Planeta Actimedia, S.A., el centro tecnológico del grupo Planeta. Se trata del prototipo de un corrector orto tipográfico y gramatical que ha sido creado dentro del marco del proyecto SCASEM (Sistema de Catalogación Semántica), cuyo objetivo es la introducción de la tecnología del procesamiento del lenguaje natural dentro del entorno editorial del Grupo Planeta.

1. El Corrector de Planeta

El Corrector de Planeta pretende ser el corrector orto tipográfico y gramatical más completo y eficaz del mercado mediante una integración estratégica de distintos módulos, cada uno de los cuales podría constituir una herramienta equiparable a las actualmente comercializadas como un producto completo.

Esto le dará la capacidad, hasta ahora desconocida, de ser aplicado en el mundo editorial, una industria en la cual, por la complejidad de las tareas de corrección que allí se llevan a cabo, los correctores automáticos conocidos hasta la fecha nunca han logrado penetrar. El Corrector de Planeta tiene, pues, el objetivo de llegar a ser una herramienta de uso común en el entorno editorial, funcionando como herramienta indispensable de apoyo a los redactores y editores.

Mientras que la detección de errores ortográficos es relativamente simple y general, la detección de errores gramaticales depende en alto grado del tipo de texto que se está tratando (texto descriptivo, novela, manual técnico, etc.), y del rigor que el redactor quiera aplicar a la hora de corregirlo.

Para que los tipos de errores que se detecten sean los más adecuados a las necesidades específicas de los editores y redactores, se ha compilado una lista, más o menos exhaustiva, de los errores con los que se suelen encontrar. Partiendo de esta lista se ha creado una gramática general del castellano que incluye cláusulas de detección y, en la medida de que sea deseable, de corrección automática de los errores detectados.

2. La arquitectura del Corrector de Planeta

La arquitectura del Corrector de Planeta, desde el punto de vista de sus funciona-lidades, puede distribuirse entre los siguientes módulos:

a) Fenómenos superficiales y tipo-grafía: signos de puntuación, espaciado, mayúsculas, minúsculas... (Autómata de corrección superficial o Morfografía)

b) Parentización

c) Corrección ortográfica

d) Corrección gramatical y estilística por reglas de patrones (Autómata de corrección estructural)

e) Corrección gramatical general sobre una gramática sintagmática

Desde el punto de vista del flujo de datos, el procesamiento completo de un texto puede sintetizarse en la siguiente secuencia:
1) Análisis de parentización; 2) análisis de errores superficiales y puntuación con consulta a un diccionario de abreviaturas; 3) análisis de la ortografía con acceso a un diccionario de formas y morfología flexiva y derivativa; 4) búsqueda de candidatos para las palabras no reconocidas; 5) análisis morfológico y lematización con etiquetación mediante un 'tagger' markoviano; 6) carga de la información léxica; 7) detección de errores mediante patrones; 8) análisis gramatical lo más completo posible mediante un analizador tabular ascendente; y 9) experto en la gestión de errores gramaticales.

El módulo de detección de errores con reglas por patrones permite obtener una gran eficacia en el tratamiento de errores comunes, que pueden corregirse de manera determinista sobre contextos muy limitados. También está abierto a una fácil edición de reglas y a su

parametrización según requerimientos de control estilístico.

El módulo de corrección gramatical general se sustenta sobre una gramática sintagmática muy completa del español estándar. En sus reglas se han flexibilizado los requerimientos de gestión de rasgos para poder dar cobertura a través de la ponderación de análisis alternativos a un conjunto importante de errores en español. Se consigue así tratar correctamente problemas para los que es necesario un riguroso y completo conocimiento léxico y gramatical.

3. Ejemplos de Fenómenos tratados según los módulos

3.1 Autómata de corrección superficial

- Destacado de palabras
  "*hacer novillos*" → "hacer novillos"
- Espaciado
  'veranos ,' →'veranos,'
- Secuencia errónea de signos
  '¿Quién es?.' →'¿Quién es?'
- Corte incorrecto en final de línea
  '1\n%' →'\n1%'

3.2 Parentización

- Parentización desequilibrada
  '(... de ejemplares.' →'(... de ejemplares).'
- Signos de entonación desequilibrados
  'Por qué ...?' → '¿Por qué ...?'
- Signos delimitadores desequilibrados
  "un gran pastel... → "un gran pastel ..."
- Entrecomillado asimétrico
  '<<palabrería<<' →'<<palabrería>>'

3.3 Corrección ortográfica

- Ortografía de nombres propios
  'New Delhi' →'Nueva Delhi'
- Ortografía de nombres comunes
  'ingerencia' →'injerencia'
- Nombres propios en minúsculas
  'buda' →'Buda'
- Partes de nombres propios
  'la Habana' →'La Habana'

3.4 Autómata de corrección estructural

- Expresiones
  'Había mogollón de…' →'Había una gran cantidad de…'
- Violación de reglas ortográficas suprasintagmáticas
  'mujeres o hombres' →'mujeres u hombres'
- Minúsculas tras signo de fin de frase
  'infructuosos. también' → 'infructuosos. También'
- Falta de punto final
  'Urdu (Pakistán)\n', 'Urdu (Pakistán).\n'
- Secuencia incorrecta de signos
  '1000 km^2.,' →'1000 km^2,'

3.5 Corrección gramatical general

- Falta de concordancia nominal
  'todos lo días' →'todos los días'
- Uso de ' que' por ' qué'
  '¿En que ...?' →'¿En qué ...?'
- Omisión de nexo sintáctico
  '2.000.000 habitantes' →'2.000.000 de habitantes'
- Falta de concordancia verbal
  'Las fachadas de las casas recuerdas las casas' →
  'Las fachadas de las casas recuerdan las casas'
- Adición de nexo sintáctico
  'la revolución convirtió a Moscú' -> 'la revolución convirtió Moscú'
- Estructura oracional incorrecta
  'no agotar el fondo las especies marinas' →'no agotar las especies marinas'
- Estructura sintáctica no oracional incorrecta
  'le primer líder' →'el primer líder'