AGME: Un Sistema de Análisis y Generación de la Morfología del Español

Francisco Velásquez, Alexander Gelbukh, Grigori Sidorov

Centro de Investigación en Computación (CIC),
Instituto Politécnico Nacional (IPN),
Av. Juan de Dios Bátiz, esq. con Miguel Othón de Mendizábal,
México D. F., Zacatenco, CP 07738, México
fcastillov@ipn.mx, {gelbukh, sidorov}@cic.ipn.mx

Resumen

La mayoría de los sistemas de análisis morfológico están basados en un modelo conocido de dos niveles. Sin embargo, este modelo no es muy adecuado para lenguajes con alternaciones irregulares de raíz (por ejemplo, el español o el ruso). En este artículo describimos un sistema computacional de análisis morfológico para el lenguaje español basado en otro modelo, cuya idea principal es el análisis a través de generación. El modelo consiste en un conjunto de reglas para obtener todas las raíces de una forma de palabra para cada lexema, su almacenamiento en el diccionario, la producción de todas las hipótesis posibles durante el análisis y su comprobación a través de la generación morfológica. Se usó un diccionario de 40,000 lemas, a través del cual se pueden analizar más de 2,500,000 formas gramáticas posibles. Para el tratamiento de palabras desconocidas se está desarrollando un algoritmo basado en heurísticas. El sistema desarrollado está disponible sin costo para el uso académico.

Palabras clave: análisis morfológico automático, generación morfológica automática, análisis a través de generación, español.

Abstract

Most widely spread method that is used in systems of automatic morphological analysis is based on a well-known two-level model. Still this model is not well suit for languages with irregular stem alternations (like, e.g., Spanish or Russian). In this paper we describe a system with automatic morphological analysis for Spanish based on other model, basic idea of which is analysis through generation. This model consists in a set of rules that allow for obtaining of all possible stems for each lexeme, their storage in the dictionary, producing of all possible hypotheses during analysis, and verification of hypotheses through morphological generation. We used a dictionary containing 40.000 lemmas, using which more than 2.500.000 possible grammatical forms can be recognized. For treatment of unknown words a heuristic-based algorithm is being developed. The system is freely available for academic use.

Key words: automatic morphological analysis, automatic morphological generation, analysis through generation, Spanish.

1. Introducción

La morfología estudia la estructura de las palabras y su relación con las categorías gramáticas del lenguaje. El objetivo del análisis morfológico automático es llevar a cabo una clasificación morfológica de una forma de palabra. Por ejemplo, el análisis de la forma *gatos* resulta en *gato +Noun+Masc+Pl*, que nos indica que se trata de un sustantivo plural con género masculino y su forma normalizada (lema) es *gato*.

Muchos procesadores morfológicos están basados en el modelo de dos niveles de Kimmo Koskenniemi [Koskenniemi, 1983]. Originalmente, el modelo fue desarrollado para el lenguaje finlandés, después se le hicieron algunas modificaciones para diferentes lenguajes (inglés, árabe, etc.). Como modelo computacional general, tiene la ventaja de que permite describir la morfología del lenguaje como un conjunto de reglas para un autómata finito. Muy poco después de la disertación de Koskenniemi, Lauri Karttunen y otras personas desarrollaron una implementación en LISP del modelo de dos niveles y lo nombraron PC-KIMMO. Sin embargo, la complejidad del sistema morfológico para la tarea de análisis automático no depende tanto del número de las clases gramaticales ni de la homonimia de flexiones, sino del número de las alternaciones en las raíces, las cuales no se pueden deducir sin consultar el diccionario, por ejemplo, *mover-muevo* vs. *correr-corro*. Para el caso del finlandés, hay muchas alternaciones, pero la gran mayoría son deducibles sin el uso del diccionario. En general, el modelo de dos niveles es difícil de aplicar para los lenguajes donde se presenta este tipo de complejidad, como por ejemplo, el español, o en el mayor grado, el ruso ([Bider and Bolshakov, 1976]) u otras lenguas eslavas.

Hay otros modelos de análisis morfológico. Para el español, [Moreno and Goñi, 1995] proponen un modelo para el tratamiento completo de la flexión de verbos, sustantivos y adjetivos. Este modelo está basado en la unificación de características y depende de un léxico de alomorfos tanto para raíces y flexiones. Las formas de palabras son construidas por la concatenación de alomorfos por medio de características contextuales especiales. Se hace uso de gramáticas de cláusulas definidas (DCG) incluidas en la mayoría de las implementaciones en Prolog. En este modelo no se implementó un diccionario grande (se necesita mucho esfuerzo, porque hay que hacerlo manualmente). Tampoco Prolog es un lenguaje muy eficiente y presenta ciertas dificultades relacionadas con el desarrollo de interfaces usando otros lenguajes de programación.

Otro ejemplo es la clasificación de los métodos de análisis morfológico propuesta por R. Hausser [Hausser, 1999], donde los métodos se clasifican en basados en formas, basados en morfemas y basados en alomorfos. Pero el autor no propone ningún método para el análisis de las formas con alternaciones en ninguna de las tres clases (por ejemplo, la verificación de qué alomorfo debe formarla), y no se describe la implementación alguna de su modelo.

En un extremo, podemos almacenar todas las formas gramáticas en un diccionario, con el lema y toda la información gramática asociada con la forma. Con esta aproximación, un sistema morfológico es solo una gran base de datos de pocas columnas. Esto es posible para lenguajes flexivos (aunque no para aglutinativos o polisintéticos). Las computadoras modernas tienen la posibilidad de almacenar bases de datos con toda la información gramática para grandes diccionarios de lenguajes flexivos (un aproximado de 20 a 50 megabytes para el español o el ruso). Sin embargo, tales modelos tienen sus desventajas, por ejemplo, no permitirán el procesamiento de palabras desconocidas.

En teoría, ya que la morfología de cualquier lenguaje flexivo es finita, cualquier método de análisis basado en diccionario da resultados igualmente correctos. Sin embargo, no todos los métodos son igualmente convenientes de usar y fáciles de implementar.

La razón de la diversidad de los modelos es que diferentes lenguajes tienen diferente estructura morfológica; los métodos apropiados para lenguajes con morfología pobre (como el inglés) no son los mejores para los lenguajes flexivos (como el español o el ruso).

Un aspecto crucial en el desarrollo de un sistema de análisis es el tratamiento de raíces alternas regulares (*deduc-ir – deduzc-o*). El procesamiento explícito de tales variantes en el algoritmo es posible pero requiere desarrollo de muchos modelos y algoritmos adicionales, que ni son intuitivamente claros ni fáciles de desarrollar. Para la solución de esta problemática, nuestro sistema implementa el modelo desarrollado en [Gelbukh and Sidorov, 2002] que consiste en preparación de las hipótesis durante el análisis y su verificación usando un conjunto de reglas de generación. La ventaja del modelo de análisis a través de la generación son la simplicidad y la facilidad de implementación. El sistema AGME (Analizador y Generador de la Morfología del Español) presentada este artículo implementa dicho modelo.

En el resto del artículo primero se presentan las consideraciones con el modelo morfológico del español para el análisis automático; después se presenta el método de generación y análisis que se usó en el sistema AGME, es decir, el procedimiento de preparación de los datos, algoritmo de generación y algoritmo de análisis; finalmente, se presenta brevemente la implementación del sistema y se dan las conclusiones.

2. Consideraciones con el Modelo de la Morfología del Español para el Procesamiento Automático

La morfología del español no es materia trivial. Como lenguaje flexivo, el español muestra una gran variedad de procesos morfológicos, particularmente los no concatenativos. Algunos de los problemas que se presentan en un procesador morfológico del español, a decir de [Moreno and Goñi, 1995], son:

- Un paradigma verbal muy complejo. Para tiempos simples, hay alrededor de 61 formas flexivas, incluyendo el duplicado subjuntivo pasado imperfecto (6 formas). Si agregamos las 45 posibles formas para tiempos compuestos, hay 112 formas flexivas posibles para cada verbo. Pero en nuestro caso ignoramos los tiempos compuestos, porque cada cadena de caracteres en nuestro nivel se procesa por separado.
- La frecuente irregularidad de raíces y terminaciones verbales. Verbos muy comunes, como *tener*, *poner*, *poder*, *hacer*, etc., tienen hasta 7 raíces: *hac-er*, *hag-o*, *hic-e*, *ha-ré*, *hi-zo*, *haz*, *hech-o*.
- Huecos en algunos paradigmas verbales. En los llamados verbos defectivos algunas formas se pierden o simplemente no se usan. Por ejemplo, los verbos meteorológicos como *llover*, *nevar*, etc., son conjugados sólo en tercera persona del singular. Otros son más peculiares, como *abolir* que falla en primera, segunda y tercera persona del singular y tercera del plural del presente de indicativo, en presente del subjuntivo y en la segunda persona del singular de la forma imperativa. En otros verbos, los tiempos compuestos se excluyen del paradigma, como en *soler*.
- Participios pasados duplicados. Una cantidad de verbos tienen dos formas alternas, ambas correctas, como *impreso*, *imprimido*. En tales casos, el análisis debe tratar las dos formas como correctas.

Existen algunos verbos altamente irregulares que pueden ser manejados sólo al incluir sus formas directamente en el diccionario (como *ser*, *haber*, etc.).

Algunos sustantivos y adjetivos presentan formas alternativas correctas para el plural (ej. *bambú*, *bambús*, *bambúes*).

Hay un pequeño grupo (3%) de sustantivos invariantes con la misma forma para el singular y el plural (ej. *crisis*). Por otro lado, 30% de los adjetivos presentan la misma forma para el masculino y el femenino (ej. *azul*). Existen también los s*ingularia tantum*, donde sólo se usa la forma singular, como en *estrés*; y los p*luralia tantum*, donde sólo se usa la forma de plural, como en *matemáticas.*

A diferencia de la morfología verbal, los procesos nominales no producen cambios internos en la raíz causado por la adición de un sufijo de género o plural, a pesar de que puede haber muchos alomorfos producidos por cambios de ortografía (*luz*, *luc-es*). Obviamente, para el sistema de análisis automático se tratan como raíces diferentes.

Todos estos fenómenos sugieren que no hay un modelo simple (unificado como el modelo de dos niveles) para el tratamiento automático de la morfología del español.

3. Modelos usados

En el español, los procesos flexivos ocurren principalmente en los nombres (sustantivos y adjetivos) y verbos. Las demás categorías gramaticales (adverbios, signos de puntuación, conjunciones, preposiciones, etc.), presentan poca o nula alteración flexiva. El tratamiento de estas últimas se realiza mediante la consulta directa al diccionario.

3.1. Morfología Nominal

La variedad de designaciones a que aluden los dos géneros y la arbitrariedad en muchos casos de la asignación de masculino o femenino a los significados de los sustantivos impiden determinar con exactitud lo que significa realmente el género. Es preferible considerarlo como un accidente que clasifica los sustantivos en dos categorías combinatorias diferentes, sin que los términos masculino o femenino prejuzguen ningún tipo de sentido concreto [Llorac, 2000].

No existen reglas estándar para la flexión de género en sustantivos. Por lo tanto, en nuestro contexto, se

almacenan todas las formas de sustantivos singulares en el diccionario (ej., *gato* y *gata*). Los adjetivos siempre tienen ambos géneros, de tal manera que solo una raíz se almacena en el diccionario, por ejemplo, *bonit-* para *bonito* y *bonita*. Ahora bien, el tratamiento de la flexión de número puede ser modelado mediante un conjunto de reglas.

Por ejemplo, las formas nominales terminadas en consonante que no sea /*s*/, agregan *–es* en su pluralización (ej., *árbol*, *árboles*). Por otra parte, los nombres acabados en vocal *–á*, *–í*, *–ó*, *–ú* tienden a presentar un doble plural en *–s* y *–es* (*esquí*, *esquíes*; *tabú*, *tabúes*), aunque algunos de ellos sólo admiten *–s* (*mamá*, *dominós*, *papás*, etc.). La información de plural no estándar se representa a través de las marcas en el diccionario para las raíces correspondientes.

3.2. Morfología Verbal

Clasificamos a los verbos en regulares (no presentan variación de raíz, como *cantar*), semi-irregulares (no más de cuatro alomorfos de raíces, como *buscar*) e irregulares (más de cuatro variantes de raíz, como *ser*, *estar*).

Afortunadamente, la mayoría de los verbos en español (85%) son regulares. Para estos, usamos los tres modelos de conjugación tradicionales (representados, por ejemplo, con los verbos *cantar*, *correr* y *partir*).

Se usan doce modelos de conjugación verbal diferentes para verbos semi-irregulares. Cada modelo tiene su tipo de alternación y su plantilla de raíces. Por ejemplo, en el modelo A1 se encuentra el verbo *buscar* (entre otros). Tiene dos raíces posibles, en este caso *busc–*, *busqu–*; la segunda raíz se usa para todas las personas del presente de subjuntivo, primera persona del singular del pretérito indefinido de indicativo y en algunos casos del imperativo; la primera raíz se usa en todos los demás modos y personas.

Se usó una plantilla (cadena de números) para cada modelo de conjugación semi-irregular. Cada posición representa una conjugación posible (por ejemplo, la primera posición representa la primera persona del singular del presente de indicativo); las últimas posiciones hacen referencia a las formas no personales. Los números usados en la plantilla van del 0 al 4, en donde 0 indica que no hay forma posible; 1 es la raíz original; 2, 3 y 4 son las demás raíces posibles. Por ejemplo, para el modelo A1 se tiene la siguiente plantilla:

1111111111112111111111111111222222111111
1111111111111221

Esta estructura nos facilita el proceso de generación de formas verbales. Nótese que son 61 posibles formas, ya que no tomamos en cuenta las formas verbales compuestas (como, por ejemplo, *haber buscado*) porque cada parte se procesa por separado.

Al ser mínimo el número de verbos completamente irregulares (como *ser*, *estar*, *haber*), su tratamiento consistió en almacenar todas las formas posibles en el diccionario. El proceso de análisis para estas palabras consiste generar la hipótesis de un verbo irregular con la flexión cero, la cual se verifica a través de generación. El proceso de generación en este caso consiste en buscar tal palabra en el diccionario, obtener todas sus variantes y desplegar el campo de la información.

4. Preparación de los Datos

Necesitamos de la preparación preliminar de datos que consistió de los siguientes pasos principales:

- Describir y clasificar todas las palabras del lenguaje (español) en clases gramaticales y las marcas adicionales, como, por ejemplo, *pluralia tantum* (esta información se tomó de los diccionarios existentes sin inventar algunas clases adicionales);
- Convertir la información léxica disponible en un diccionario de raíces (sólo la primera raíz necesita ser generada en este paso);
- Aplicar los algoritmos de generación de raíces para generar todas las raíces (con duplicación de información de la primera raíz más el número de la raíz generada).

Se diseñó una estructura de almacenamiento de datos como la que se muestra en la Tabla 1. Para los datos mostrados, el campo *Word* contiene el lema, el campo *Stem* contiene la raíz, el campo *Info* contiene la clase gramatical, los campos *Mark1, Mark2* contienen las marcas gramaticales adicionales. Por ejemplo, el campo *Mark1* del registro 2 (*P*) indica que se trata de un *pluralia tantum* (es decir, si se tratara de generar su singular, obtendríamos un error) y para los últimos dos registros indica el modelo de conjugación semi-irregular al que pertenece el verbo. El campo *Mark2* para los últimos dos registros señala la raíz original (*1*) y la segunda raíz posible (*2*).

Tabla 1. Estructura del diccionario de raíces.

Stem	Word	Info	Mark1	Mark2
gato	gato	N		
gafa	gafas	N	P	
acert	acertar	VI	M1	1
aciert	acertar	VI	M1	2

5. Proceso de Generación

El proceso de generación se desarrolla de la siguiente manera. Tiene como entrada los valores gramaticales de la forma deseada y la cadena que identifique la palabra (cualquiera de las posibles raíces o el lema).

- Se extrae la información necesaria del diccionario;
- Se escoge el número de la raíz necesaria según las plantillas;
- Se genera la raíz necesaria;
- Se elige la flexión correcta según el algoritmo desarrollado (el algoritmo es bastante simple y obvio, por ejemplo, para el verbo de clase 1 en primera persona, plural, indicativo presente la flexión es *-amos*, etc.), y
- La flexión se concatena con la raíz.

6. Proceso de Análisis

El modelo general de análisis morfológico mostrado en Fig. 1 e implementado en nuestra aplicación, es simple: dependiendo de la forma de palabra de entrada, se formula alguna hipótesis de acuerdo con la información del diccionario y otros criterios y se generan las formas correspondiente para tal(es) hipótesis. Por ejemplo, para la flexión *-amos* y la información del diccionario para la raíz que corresponde al verbo de la clase 1, se genera la hipótesis de primera persona, plural, indicativo presente (entre otras), etc.

Las formas generadas según las hipótesis se comparan con la original, en caso de coincidencia las hipótesis son correctas.

Más detalladamente, dada una cadena de letras (forma de palabra), la analizamos de la siguiente manera:

1. Quitar letra por letra (también siempre se verifica la hipótesis de la flexión ∅).
2. Verificar si existe flexión.
3. Si existe flexión entonces leer del diccionario la información de la raíz y llenar la estructura de datos correspondiente (si no existe la raíz, regresar al paso 1).
4. Si no existe la flexión, regresar al paso 1.
5. Formular hipótesis.
6. Generar la correspondiente forma gramatical de acuerdo a nuestra hipótesis y la información del diccionario.
7. Si el resultado obtenido coincide con la forma de entrada entonces la hipótesis es aceptada. De otra forma, el proceso se repite desde el paso 3 con otra raíz homónima (si la hay) o desde el paso 1 con otra hipótesis sobre la flexión.

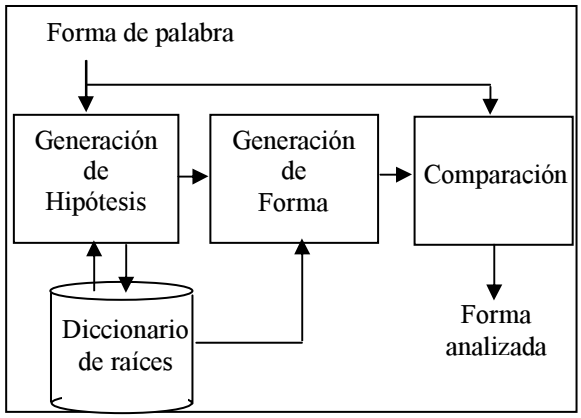

Fig. 1. Proceso de análisis morfológico.

Nótese que es importante la generación porque de otro modo algunas formas incorrectas se aceptarían por el sistema, por ejemplo, **acuerdamos* (en lugar de *acordamos*). En este caso existe la flexión y existe la raíz, pero son incompatibles, lo que se verifica a través de la generación.

7. Implementación

La base de datos (el diccionario) es una tabla Paradox donde se almacenan las raíces y otra información de las mismas, como se mostró en el apartado de preparación de los datos (Sección 4).

El sistema se codificó en C++. Cuenta con una interfaz que permite escoger la forma gramatical para generar o para introducir la palabra para el análisis. También existe una versión del sistema, la cual lee un archivo de texto grande, en vez de procesar una sola forma de palabra.

8. Conclusiones

Se presentó un sistema de análisis morfológico que implementa el modelo de comprobación de hipótesis a través de generación. Las ventajas de este modelo de análisis reflejadas en su implementación son su simplicidad, la velocidad en la que se implementó y la claridad. El desarrollo de los algoritmos principales sólo tomó varios días.

El diccionario actual tiene un tamaño considerable: 40,000 lemas, incluyendo 23,400 sustantivos, 7,600 verbos y 9,000 adjetivos.

Es importante mencionar que AGME no sobre-genera o sobre-analiza, es decir, sólo se procesan las formas correctas.

Se está trabajando en el tratamiento de enclíticos y en la forma de tratar las palabras desconocidas (una aproximación inicial es la de formular una heurística del más parecido). Como trabajo futuro se sugiere centrar los esfuerzos a los procesos de derivación (*bella* ⇒ *belleza*) y composición (*agua* + *fiesta* ⇒ *aguafiestas*).

El sistema desarrollado está disponible como archivo EXE o DLL de Windows, sin costo alguno para uso académico.

Agradecimientos

Este trabajo fue realizado con el apoyo parcial del gobierno de México (CONACyT, SNI), del Instituto Politécnico Nacional, México (CGEPI-IPN, COFAA, PIFI) y la red RITOS-2 del Subprograma VII de CYTED.

Referencias

[Bider and. Bolshakov, 1976] Bider, I. G. and I. A. Bolshakov. Formalization of the morphologic component of the Meaning - Text Model. 1. Basic concepts (in Russian with a separate translation to English). ENG. CYBER. R., No. 6, 1976, p. 42-57.

[Gelbukh and Sidorov, 2002] Gelbukh, A. and G. Sidorov. Morphological Analysis of Inflective Languages through Generation. Revista *Procesamiento de lenguaje natural*, España, vol. 29, 2002, pp 105-112. (2002)

[González and Vigil, 1999] González, B. M. y C. Ll. Vigil. *Los Verbos Españoles*. 3ª Edición, España, Ediciones Colegio de España. 1999. 258 p. (1999)

[Hausser, 1999] Hausser, R. *Three Principled Methods of Automatic Word Form Recognition*. Proc. of VEXTAL: Venecia per il Tratamento Automatico delle Lingue. Venice, Italy, 1999. pp. 91-100. (1999).

[Koskenniemi, 1983] Koskenniemi, K. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Tesis Doctoral. Universidad de Helsinki. 1983. 160 p. (1983).

[Llorac, 2000] Llorac, E. *Gramática de la Lengua Española*. España, Ed. Espasa, 2000. 406 p. (2000).

[Moreno and Goñi, 1995] Moreno, A. and J. Goñi. *GRAMPAL: A Morphological Processor for Spanish Implemented in PROLOG*. En: Mar Sessa y María Alpuente, editores, Proceedings of the Joint Conference on Declarative Programming (GULP-PRODE'95), pp. 321-331, Marina di Vietri (Italia), 1995. (1995)

[Santana *et al.*, 1999] Santana, O., J. Pérez, *et al*. *FLANOM: Flexionador y Lematizador Automático de Formas Nominales*. Universidad de las Palmas de Gran Canaria. Lingüística Española Actual XXI, 2. Ed. Arco/Libros, S.L. España, 1999. (1999)