Introducción al Procesamiento de Lenguaje Natural

Grupo PLN – InCo

Morfología

Morfología

morfología (de *morfo* (forma) y *logía* (ciencia))

Rama de una disciplina que se ocupa del **estudio y la descripción de las formas externas de un objeto**

Se puede aplicar al estudio de:

- los seres vivos (Biología)
- la superficie terrestre (Geomorfología)
- las palabras (Lingüística)

Morfología

f. *Gram.* Parte de la lingüística que se ocupa de la estructura o forma de las palabras.

Morfología

f. *Gram.* Parte de la lingüística que se ocupa de la estructura o forma de las palabras.

"La morfología explica la estructura interna de las palabras y el proceso de formación de palabras mientras que la sintaxis describe cómo las palabras se combinan para formar sintagmas, oraciones y frases."
(Wikipedia)

Morfología

- Mecanismos de formación / análisis de la palabras
- **Análisis morfológico**: Reconocer una palabra y construir una representación estructurada

Gatitos → gato + Masc + Pl + Dim

Morfología

- Texto: describe a un conjunto de oraciones que permite dar un mensaje coherente y ordenado de manera escrita

- Oración: constituyente sintáctico más pequeño necesario para expresar un predicado completo

- Enunciado: la versión para lenguaje hablado de la oración

Morfología

- Morfema: fragmento mínimo capaz de expresar significado. Es la mínima unidad con sentido
- Raíz: es la parte de la palabra que no varía y que indica su significado principal. Es el "morfema "principal" o lexema

Los morfemas se añaden a la raíz para formar nuevas palabras.

Morfología

- Morfema: fragmento mínimo capaz de expresar significado. Es la mínima unidad con sentido
- Raíz: es la parte de la palabra que no varía y que indica su significado principal. Es el "morfema "principal" o lexema

Los morfemas se añaden a la raíz para formar nuevas palabras.

- Afijos: dan significado adicional

Morfología

- Morfema: fragmento mínimo capaz de expresar significado. Es la mínima unidad con sentido
- Raíz: es la parte de la palabra que no varía y que indica su significado principal. Es el "morfema "principal" o lexema

Los morfemas se añaden a la raíz para formar nuevas palabras.

- Afijos: dan significado adicional
  - Prefijos: im+posible

Morfología

- Morfema: fragmento mínimo capaz de expresar significado. Es la mínima unidad con sentido
- Raíz: es la parte de la palabra que no varía y que indica su significado principal. Es el "morfema "principal" o lexema

Los morfemas se añaden a la raíz para formar nuevas palabras.

- Afijos: dan significado adicional
  - Prefijos: im+posible
  - Sufijos: gat+ito+s

Morfología

- Morfema: fragmento mínimo capaz de expresar significado. Es la mínima unidad con sentido
- Raíz: es la parte de la palabra que no varía y que indica su significado principal. Es el "morfema "principal" o lexema

Los morfemas se añaden a la raíz para formar nuevas palabras.

- Afijos: dan significado adicional
  - Prefijos: im+posible
  - Sufijos: gat+ito+s
  - Circunfijos: a+naranj+ado (Parasintéticas)

Morfología

- Morfema: fragmento mínimo capaz de expresar significado. Es la mínima unidad con sentido
- Raíz: es la parte de la palabra que no varía y que indica su significado principal. Es el "morfema "principal" o lexema

Los morfemas se añaden a la raíz para formar nuevas palabras.

- Afijos: dan significado adicional
  - Prefijos: im+posible
  - Sufijos: gat+ito+s
  - Circunfijos: a+naranj+ado (Parasintéticas)
  - Infijos: No hay (hingi => h**um**ingi Lengua Tagaloga)

Morfología

- Lematización: llevar palabras con la misma raíz a una forma canónica. Identificar su estructura interna

  Por ejemplo

  en español:

  soy, son, es → ser

  perro, perra, perros → perro

  en ingles:

  am, are, is → be

- Lema: palabra “representativa”

Morfología

- Stemming: cortar las palabras. Mucho más simple, en los hechos ha funcionado

- Stemmer de Porter (1980): una serie de reglas de reescritura en cascada

Morfología

Algoritmo de Porter (ejemplo de reglas)

Step 1a

sses → ss	caresses	→	caress
ies → i	ponies	→	poni
ss → ss	caress	→	caress
s → ø	cats	→	cat

Step 1b

(*v*)ing → ø	walking	→	walk
	sing	→	sing
(*v*)ed → ø	plastered	→	plaster
…			

Step 2 (for long stems)

ational → ate	relational	→	relate
izer → ize	digitizer	→	digitize
ator → ate	operator	→	operate
…			

Step 3 (for longer stems)

al → ø	revival	→	reviv
able → ø	adjustable	→	adjust
ate → ø	activate	→	activ
…			

Morfología

Diferencia entre palabras y morfemas

- Aislabilidad – una palabra puede aparecer sola
- Permutabilidad – una palabra puede cambiar el orden en que aparece
- Función sintáctica – No puedo hacer oraciones con morfemas
- Combinación – Puedo combinar afijos, sufijos.... pero ojo

Morfología

Dos problemas a resolver:

1) Formación de las palabras (Morfotáctica)
   - Las palabras están compuestas por unidades menores (morfemas)
   - Los morfemas pueden combinarse de acuerdo a ciertas reglas

     inevitable

     *inelefante

     inelefantemente ?
   - Los morfemas NO pueden aparecer solos (en general)

Morfología

Dos problemas a resolver:

2) Alteraciones Ortográficas

Los morfemas pueden cambiar según el contexto

Pez → Pezs → Peces

Maní → Manís → Maníes

Morfología

Formas de combinar morfemas:

- Flexión
- Derivación
- Composición
- Clitización

Morfología

Morfología Flexiva

- Mecanismo de producción de palabras dentro de una misma clase

  com – o / com – ía / com – eré

- En español no se agrega significado extra

- Las flexiones aportan información relativa a:

  Género / Número Persona / Tiempo / Modo

Ejemplo: etiquetas Eagle

Morfología

Etiquetas Eagle

son un estándar para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas

Forma	Lema	Etiqueta
alegres	alegre	AQ0CS00
hábilmente	hábil	RG000
el	el	TDMS0
ninguna	ninguno	DI3FS00
ningunos	ninguno	DI3MP00
Juan	juan	NP00000
es	ser	VAIP3S0

Morfología

Morfología Derivativa (o léxica)

- Combinar una raíz con un afijo, para generar una palabra de *otra* clase, o con *otro* significado
  - descubrir (verbo) → descubrimiento (sustantivo)
  - estable (adjetivo) → estabilizar (verbo)
    → estabilización (sustantivo)
    → desestabilización (sustantivo)
- Es un mecanismo *productivo*

Morfología

Apenas él le amalaba el noema, a ella se le agolpaba el clémiso y caían en hidromurias, en salvajes ambonios, en sustalos exasperantes. Cada vez que él procuraba **relamar** las incopelusas, se enredaba en un grimado quejumbroso y tenía que **envulsionarse** de cara al nóvalo, sintiendo cómo poco a poco las arnillas se espejunaban, se iban **apeltronando**, **reduplimiendo**, hasta quedar tendido como el trimalciato de ergomanina al que se le han dejado caer unas fílulas de cariaconcia. Y sin embargo era apenas el principio, porque en un momento dado ella se **tordulaba** los hurgalios, consintiendo en que él aproximara suavemente sus orfelunios.

(*Rayuela – Julio Cortázar*)

Morfología

Tipos de lenguajes según su morfología:

- Flexionales: inglés, español
- Aislantes: chino (solamente un morfema por palabra)
- Aglutinantes: turco (pegan muchos morfemas)

  Ejemplo:
  uygarlık edemeyenler arasında olsaydınız

- Polisintéticos: lenguas indígenas americanas; esquimales (jupik)

Morfología

Tipos de lenguajes según su morfología:

- Flexionales: inglés, español
- Aislantes: chino (solamente un morfema por palabra)
- Aglutinantes: turco (pegan muchos morfemas)

  Ejemplo:

  uygarlık edemeyenler arasında olsaydınız

  comportándote como si estuvieras entre aquellos que no pudimos civilizar
- Polisintéticos: lenguas indígenas americanas; esquimales (jupik)

Morfología computacional

- Parsing (análisis): tomar una entrada y devolver algún tipo de estructura lingüística
- Parsing Morfológico: tomar una palabra (forma de superficie) y devolver una representación de la raíz y sus morfemas.

gatito <--> *gato* + NCMSooD

Nombre Común. Masculino. Singular. Diminutivo

Morfología computacional

Métodos:

1. Diccionario de formas flexionadas

amigo → amigo + NC + MS
amiga → amigo + NC + FS
amigas → amigo +NC + FP
amiguitos → amigo +NC + MPD
…

Morfología computacional

1. Diccionario de formas flexionadas

- Hay que armarlo
- Se vuelven enormes
- No permiten palabras nuevas

... y siempre podemos inventar

Morfología computacional

¿Qué pasa con los lenguajes aglutinantes?

taloissani

puede segmentarse de la siguiente forma:

talo → casa
i → indica que es plural
ssa → indica que está "dentro de"
ni → sufijo que indica al posesor de primera persona “mi, mis”

en mis casas

Morfología computacional

¿Qué pasa con los lenguajes aglutinantes?

- El turco tiene 40,000 formas posibles
- A eso hay que sumarle los sufijos derivativos
- Potencialmente hay infinitas palabras
- Otros lenguajes aglutinantes: aymará, euskera, finlandés

Morfología computacional

Métodos:

2. Parser morfológico dinámico

pavos → pavo +N +Masc +Pl

palabra → lema + rasgos

Necesitamos

1. Un lexicón: lista de raíces y afijos
2. Morfotácticas: modelo de ordenamiento de morfemas; que clases de morfemas pueden seguir a otros
3. Reglas ortográficas: cambios en una palabra al combinar morfemas

Morfología de Estado Finito

Morfología de Estado Finito

Desde 1968...

Las alteraciones o cambios fonéticos se describían por reglas de reescritura (Chomsky y Halle) ordenadas

$\alpha \rightarrow \beta \ / \ \gamma \_ \delta$

α reescribe en β si aparece entre γ y δ

n → m / i _ p [i**n**posible → i**m**posible]

Pero no se sabía como usarlas para el análisis

Son reglas de reescritura contextuales... tienen un poder expresivo igual a una Máquina de Turing

Morfología de Estado Finito

1972

Johnson: los fonologistas siempre asumieron que luego del reemplazo, nos movemos en el string, no vuelve a aplicarse la regla

En esas hipótesis, las reglas de reescritura pueden representarse por *transductores*

Morfología de Estado Finito

aprox 1980

Kaplan & Kay: Regular Models of Phonological Rule Systems

Las reglas de reescritura describen RELACIONES REGULARES

.... que se pueden ver como transductores

Estaba toda la teoría, pero llevó mucho tiempo implementar las operaciones básicas

Morfología de Estado Finito

Transductores léxicos

- Hacen parsing morfológico
- Una palabra es una correspondencia
  - Nivel léxico (colección de morfemas)
  - Nivel de superficie

g|a|t|i|t | o| | | | |
g|a|t|o|+Masc|+Sg|+D| | | |

- Vemos a los FST como autómatas sobre dos cintas
- Dos alfabetos diferentes

Transductores de Estado Finito

Transductores de Estado Finito

Autómatas Finitos con transiciones sobre parejas de símbolos

Definición:

Un Transductor de Estado Finito (FST) es una tupla:

$T=(Q,\Sigma_1, \Sigma_2 ,i,F,E)$

Q conjunto finito de estados

Σ_1, Σ_2 alfabetos de entrada y salida

i estado inicial $\in$ Q

F conj. finito de estados finales $F \subseteq Q$

E conj. finito de aristas $E \subseteq Q \times \Sigma_1 \cup \{\varepsilon\} \times \Sigma_2 \cup \{\varepsilon\} \times Q$

Transductores de Estado Finito

Definición:

Relación Asociada a un Transductor

Sea $T=(Q,\Sigma_1, \Sigma_2, i, F, E)$

Queda definida una relación $L(T) \subseteq \Sigma 1^* \times \Sigma 2^*$

$L(T) = \{(x,y) \in \Sigma 1^* \times \Sigma 2^* / \exists (i,x,y,q) \in \hat{E}, \text{ con } q \in F\}$

$\hat{E}$ (clausura del conj. de aristas): menor conjunto tal que:

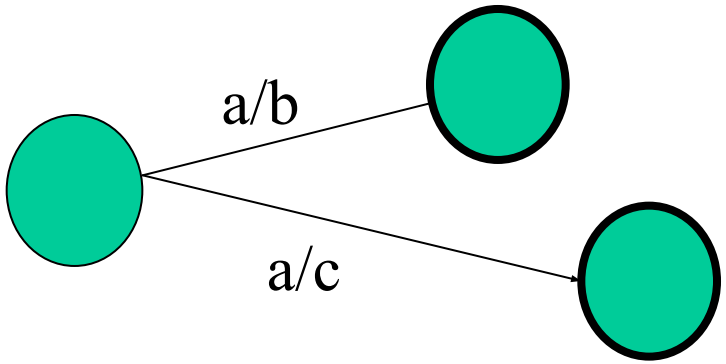
(i) $\forall q \in Q, (q,\varepsilon, \varepsilon,q) \in \hat{E}$

(ii) $\forall x \in \Sigma 1^*, \forall y \in \Sigma 2^*, \forall a \in \Sigma 1 \cup \{\varepsilon\}, \forall b \in \Sigma 2 \cup \{\varepsilon\}$, si $(q1,x,y,q2) \in \hat{E}$ y $(q2,a,b,q3) \in E$, entonces $(q1,xa,yb,q3) \in \hat{E}$

Transductores de Estado Finito

Estado Finito = Eficiencia

Los FST se pueden ver como autómatas sobre pares de símbolos, y por lo tanto determinizar, minimizar, etc.



Transductores de Estado Finito

Algunas herramientas:

- Xerox twolc: Two Level Compiler
- At&T FSM Tools: manipulación de FSTs
- FSA Utilities: manejo de expresiones regulares, autómatas. Permite definir nuevos operadores. En Prolog.
- OpenFST, la versión Open Source de FSM Tools

Referencias

- J.Martin & D.Jurafsky. Speech and Language Processing - Capítulo 3
- K.Beesley, L.Karttunen. Finite State Morphology
- Mehryar Mohri - Finite-State Transducers in Language and Speech Processing.