



Extracción de la Información

SEI-GO

31 Mayo, 2010

Inteligencia Artificial

Integrante	LU	Correo electrónico
Cecilia Sanchez	000/00	chechus26@gmail.com
Matías Pérez	002/05	elmaildematiaz@gmail.com

Reservado para la cátedra

Instancia	Corrector	Nota
Primera entrega		
Segunda entrega		



**DEPARTAMENTO
DE COMPUTACION**

Facultad de Ciencias Exactas y Naturales - UBA

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Ciudad Universitaria - (Pabellón I/Planta Baja)

Intendente Güiraldes 2160 - C1428EGA

Ciudad Autónoma de Buenos Aires - Rep. Argentina

Tel/Fax: (54 11) 4576-3359

<http://exactas.uba.ar/>

Índice general

0.1. Introduccion	II
0.1.1. Obejtivo	II
0.1.2. Alcance	II
0.2. Descripcion	III
0.2.1. Armado del diccionario	III
0.2.2. Definición de reglas	III
0.2.3. Aplicación de reglas	IV
0.2.4. Recolección de resultados	IV
0.2.5. Consideraciones finales	IV
0.3. Metodología	V
0.4. Resultados	VI
0.5. Conclusiones	VII
0.6. Trabajo Futuro	VIII
0.7. Anexos	IX
0.7.1. Boludeces de latex	IX
0.7.2. Tabla Loca	IX

0.1. Introduccion

0.1.1. Obejtivo

El trabajo consiste en la realización de un sistema de extracción de información para la *GuiaOleo*¹, para poder conseguir información relevante a partir de los comentarios que se encuentran en la misma.

El criterio usado para definir si cierta información es relevante o no es si la misma consiste en apreciaciones sobre la comida, el servicio y otros aspectos de los restaurants. De esta forma se considera información irrelevante aquella que trate sobre la persona en sí u otros aspectos que se suelen encontrar en los comentarios personales.

Vale aclarar que dicho criterio es arbitrario, ya que se hubiera podido elegir cualquier otro criterio.

0.1.2. Alcance

Lo que se intenta obtener con este sistema es una forma sencilla y rápida de conseguir las características de los restaurants sin la necesidad de leer todos los comentarios. Se trata de obtener de los comentarios los fragmentos de los mismos que cumplen con el criterio ante dicho.

Por ejemplo, del comentario:

“Es uno de los pocos restaurantes que sirve cocina italiana de verdad y actualiza su menu; la atencion, si bien llevada mayormente por chicas jovenes es sobria, amable y eficiente. Voy muy seguido y he celebrado Navidad alli. Recomiendo tanto las pizzas como los platos de cocina. Dejen lugar para los postres...valen la pena acompañados por un ristretto lavazza autentico.”

Se desea obtener *“Recomiendo tanto las pizzas como los platos de cocina”* y *“la atencion, [...] es sobria, amable y eficiente”*, ya que comentarios como *“Voy muy seguido y he celebrado Navidad alli”* no resultan relevantes ante el criterio elegido.

En este trabajo nos conformaremos con que sólo reconozca la primera oración, ya que la segunda contiene información no relevante en el medio de la oración. Dejando el reconocimiento de este tipo de oraciones para etapas posteriores y futuros refinamientos.

¹<http://www.guiaoleo.com.ar>

0.2. Descripción

El desarrollo del sistema se puede dividir en cuatro etapas: *armado del diccionario*, *definición de reglas*, *aplicación de reglas* y *recolección de resultados*. Las mismas serán descritas en más profundidades a continuación.

0.2.1. Armado del diccionario

Antes de empezara a analizar los comentarios, fué necesario definir las palabras que definen nuestro mundo y caracterizarlas, ya que por ejemplo, no cumplen la misma función gramatical *pizza* o *excelente*.

Las entidades gramaticales que se terminaron usando son:

- Entidad: Define a los sustantivos del dominio relevantes, pueden ser comidas o mesas, ambiente, etc.
- Adjetivo: Son los adjetivos calificativos, estos a su vez se dividieron en:
 - Positivo
 - Neutro
 - Negativo
- Verb: verbos
- Art: articulos
- Empty: Son todas aquellas palabras que no están definidas como ninguna otra entidad gramatical.

En este punto, hubiera sido ideal poder definir todas las palabras del idioma castellano, pero esto no es posible, por lo que se hizo fué definir las palabras más relevantes en nuestro dominio (los comentarios de la guía oleo).

Con este fin fué necesario primero determinar cuáles son estas palabras, lo que se optó por hacer fué simplemente ordenar las palabras por cantidad de apariciones en los comentarios. Con este procedimiento se logró conseguir una primera aproximación de las palabras más relevantes.

Pero dado que el tiempo con que se contaba no era suficiente para caracterizar todas las palabras, se optó por caracterizar las palabras por su raíz, esto disminuyó notablemente la cantidad de palabras, ya que para caracterizar palabras como *malo*, *mala*, *mal*, *malos*, *malas*, *males* sólo fué necesario caracterizar *mal*.

Este método si bien ayuda a disminuir la cantidad de elementos del dominio, también trae sus consecuencias, ya que palabras como *precios* y *preciosamente* tienen la misma raíz *preci*, siendo la primera una *Entidad* y la otra un *Adjetivo*. Pero esto fué algo con lo que se aceptó seguir adelante.

Una vez obtenido un diccionario con todas las raíces ordenadas por orden de aparición se comenzaron a caracterizar las más importantes manualmente. Es necesario aclarar que la caracterización de las palabras fué considerando el contexto de la *GuíaOleo*, ya que palabras como *tierno* se tomó como un adjetivo positivo y palabras como *abuelo* no se consideraron como entidades a pesar de ser sustantivos.

Luego de completar las subsiguientes etapas, una vez que se obtuvo un sistema andando con este diccionario, se decidió enriquecer el diccionario de un modo más automático. Este proceso será explicado más adelante.

0.2.2. Definición de reglas

Una vez obtenido el diccionario de palabras a usar, el siguiente paso fué representar el criterio por el cuál se definiría si una frase es relevante o no. Esto se realizó mediante el uso de expresiones regulares.

De esta manera se definieron expresiones regulares de modo tal que sólo las frases que cumplan con estas expresiones cumplan con el criterio propuesto. Esto fué hecho manualmente y las expresiones regulares se fueron creando a base de sentido común y pruebas en el conjunto de datos de entrenamiento.

Las reglas que se obtubieron finalmente fueron las siguientes:

```
(ART)* ENTITY (EMPTY)+ (ADJETIVE)+ ( y |ADJETIVE)*  
(ART)* ENTITY (VERB)+ (ADJETIVE)+ (EMPTY|ADJETIVE)*  
(ART)* ENTITY (VERB)+ (EMPTY)+ (ADJETIVE)+ (EMPTY|ADJETIVE|ART)* (ENTITY)*  
(ART)+ (ADJETIVE)+ ENTITY (EMPTY)+ (ADJETIVE)+  
(ART)+ ENTITY (ADJETIVE)+  
(no)* (VERB)* (ART)+ (ADJETIVE)+ ENTITY  
(ART)* ADJETIVE ART ENTITY  
(ART)* VERB ADJETIVE ART ENTITY  
VERB VERB ADJETIVE  
Muy ADJETIVE (ART)* ENTITY  
uno de los ADJETIVE ENTITY para comer  
no lo recomiendo  
no es para recomendar
```

Se puede ver que la mayoría de las reglas depende sólo de las entidades definidas, pero hay casos en los que la expresión de la caracterización no es suficiente y se pierden casos de gran interés, como ser “*no lo recomiendo*”. Para no perder esto simplemente se agregó una regla que busca esta frase en particular.

En este punto es justo hacer notar que los modificadores tales como *muy*, *no*, *más*, etc. no fueron caracterizados, haciendo imposible la detección de estas palabras en frases más complejas. Al igual que no se detectan frases que contengan aposiciones u otras construcciones gramáticas con relativo grado de complejidad. Estos casos quedan para futuras iteraciones de este sistema.

0.2.3. Aplicación de reglas

Una vez que se cuenta con el diccionario de las palabras ya caracterizado y con las reglas definidas, es necesario aplicar estas reglas a los comentarios. Con este fin se desarrolló un sistema informático que simplemente busca la aparición de las expresiones regulares en un comentario.

El mismo toma los diferentes archivos con cada tipo de palabras, un archivo con las reglas definidas y por último los comentarios a evaluar.

Este devuelve un archivo de texto con todos los comentarios seguidos de las frases obtenidas y un detalle de cuál fue la regla que se usó para detectar la frase.

0.2.4. Recolección de resultados

0.2.5. Consideraciones finales

0.3. Metodología

0.4. Resultados

0.5. Conclusiones

0.6. Trabajo Futuro

0.7. Anexos

0.7.1. Boludeces de latex

Setear la codificación del editor UTF-8, es la default... pero si no se ven bien lo acentos o algo es por eso.
Bla,blalb,a

Acá puedo poner un comentario. lo identa locamente :P

Escribo sin tab

Quiero enfatizar esta *palabra*

- Item 1
- Item 2
 - sub-item1
 - sub-item2
- otro item

0.7.2. Tabla Loca

Policía	Lugar	Ladrón	País
Dodero	Avión	Dedos	Alemania
Elkin	Bar	Gato	Austria
Frigerio	Barco	Hurón	Espana
Kesner	Cine	Rata	Francia
Minari	Tren	Sombra	Inglaterra