Extracción de la Informacion

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

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 carácterísticas 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 reconocimento de este tipo de oraciones para etapas posteriores y futuros refinamientos.

[1] http://www.guiaoleo.com.ar

0.2. Descripcion

El dasarrollo del sistema se puede dividir en cuatro etapas: *armado del diccionario*, *definicion de reglas*, *aplicacion de reglas* y *recoleccion de resultados*. Las mismas serás descriptas en más profundidas a continuación.

0.2.1. Armado del diccionario

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

Las entidades gramáticales 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 gramátical.

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 guia oleo).

Con este fin fué nesesario 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 nescesario 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 raices 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 obteubo 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 expresion 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 fín se desarrolló un sistema informático que simplemente busca la aparicion de las expresiones regulares en un comentario.

El mismo toma como entrada:

- los diferentes archivos con cada tipo de palabras,
- un archivo con las reglas definidas y
- los comentarios a evaluar.

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

0.2.4. Recolección de resultados

Los resultados obtenidos se plasmaron en un archivo de texto plano, en los mismos se encuentran el comentario original analizado y las reglas que aplicaron en el mismo, seguidas de el extracto que aplicó. De esta manera se puede observar con claridad el funcionamiento del sistema de extracción de información.

Por otro lado, en base a los resultados obtenidos se realizó un análisis preliminar sobre cada uno, esto fué posible gracias a que los adjetivos fueron divididos en tres clases

- Positivos,
- Negativos,
- Neutros

De esta manera ante un comentario se hizo un balance entre los adjetivos positivos y negativos que el mismo contiene para dar una idea preliminar del matiz del mismo.

Este resultado puede ser ampliamente mejorado, ya que no se vieron aspectos como la relación entre los diferentes adjetivos o el contexto en el que se encuentran los mismos ni el peso de estos. Ya que es muy distinto el peso de un adjetivo como *exelente* y un adjetivo como *bueno*, siendo los dos adjetivos positivos.

Una vez realizado esta categorización se realiza también un resumen del restaurant teniendo en cuenta todos los comentarios del mismo y el aspecto general de cada uno de ellos.

Se decidió este acercamiento, en vez de tener en cuenta el porcentaje de adjetivos positivos en el total de los comentarios, porque lo que se busca es encontrar el porcentaje de gente que estubo conforme con el restaurant, por esto es que primero se analiza cada comentario por separado y se lo clasifica aislado del resto y luego se clasifica el restaurant en base a los mismos.

Vale aclarar que este análisis de los datos es sólo una análisis superficial, ya que el mismo excede el alcance del trabajo propuesto.

0.2.5. Consideraciones finales

0.3. Metodología

0.4. Resultados

Los resultados obtenidos lograron identificar en gran medida las partes relevantes de los comentarios, a continuación se analizaran distintos ejemplos:

> *''Fue una experiencia espantosa, la pizza parecia una galletita de agua con tomate rebajado en agua unas hojitas de radicheta y unos pedacitos infimos de queso brie. de sabor horrenda,precio carisimo, tres pizzas grandes y en promedio dos coronas chicas por persona casi 500 pesos (eramos 6 personas) . les recomiendo abstenerce de ir,* **no es un buen lugar**, *no se come bien,* **el servicio es malisimo**, *te ponen la botella de corona en la mesa sin siquiera servirte medio vaso como corresponde. un asco, deberia darles verguenza tener un restaurant asi, el peor al que he ido.''*
>
> Resultado:
>
> - R1 el servicio es malisimo
> - R5 no es un buen lugar

Lo primero que se puede apreciar es que las partes identificadas como importantes realmente lo son. Veamos que *"el servicio es malisimo"* y *"no es un buen lugar"* , ambos son dos apreciaciones sobre el restaurant en sí. En cambio partes como *"les recomiendo abstenerse de ir"* , no lo es, sino que es una comentario personal. Este último tipo de frases son las que se desan obviar.

Por otro lado también se puede ver que se escapan frases que resultarían relevantes, como ser *"no se come bien"* o *"precio carísimo"* , esto se debe a la falta de presicion en las reglas. Se intentaron incluir este tipo de frases e incluir más reglas, pero muchas veces se generaba que se identifiquen como relevantes frases que no lo son, así que se optó por no identificar frases de más ante el peligro de obviar frases relevantes.

> *''Fuimos varios a almorzar y* **nos atendieron correctamente**, **la comida es muy rica** *y* **el ambiente es descontracturado** *y casual. En resumidas cuentas,* **un buen lugar** *para* **comer una rica** *pizza, si estan por la zona de Retiro.''*
>
> Resultado:
>
> - R0 comer una rica-
> - R1 el ambiente es descontracturado-
> - R2 la comida es muy rica-
> - R5 un buen lugar-
> - R8 nos atendieron correctamente-

Veamos que en este comentario se reconoció casi en su totalidad como relevante. Lo único que se dejó afuera del mismo es que queda en la zona de retiro, esto se debe a que la localización del restaurant no cumple con el criterio propuesto para el sistema.

Otra cosa relevante a ver es que se reconocío *"un buen lugar"* y *"comer una rica"* como frases relevantes por separado, esto se debe a que los conectores como *para* no se incluyeron en el diccionario de palabras.

Por otro lado se reconoció como una buena frase a *"comer una rica"* , la regla que se aplicó para identificarla es:

```
\epmh{R0}: (ART)* ENTITY (EMPTY)+ (ADJETIVE)+ (y|ADJETIVE)*;
```

Lo que sucedió aquí es que *comer* se reconoció como una entidad, esto se debe a que la raíz de comer es *com*, y esta es raíz también de: *"comida"* , y al ser está ultima una entidad tan importante en el contexto del trabajo se optó por identificarla como entidad bajo el riesgo de confundirla con el verbo *comer*.

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