Extracción de la Informacion

SEI-GO

31 Mayo, 2010 — Inteligencia Artificial

Integrante	LU	Correo electrónico
Cecilia Sánchez	380/03	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 (SEI) para la *GuiaOleo*[1], el cuál consiste en obtener 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 o información que no esté relacionada con el lugar. Como ser *"Voy siempre en navidad"* o *"Fuí con mi pareja"*.

Dado el domino del sistema, se optó por este criterio, pero en otro contexto, el mismo puede no tener sentido.

El objetivo práctico del SEI es conseguir de manera sencilla y rápida las carácterísticas relevantes de los restaurants sin leer la totalidad de los comentarios, esto es de gran utilidad si se cuenta con un gran volumen. Es imporatante destacar que si el sistema funciona correctamente la cantidad de información a analizar es mucho menor.

Para el siguiente 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."*

Los resultados esperados sería obtener las oraciones:
"Recomiendo tanto las pizzas como los platos de cocina" y *"la atencion, [..] es sobria, amable y eficiente"*.
y obviar las oraciones como *"Voy muy seguido y he celebrado Navidad alli"* ya que no expresa información relevante.

0.1.2. Alcance

En esta primera versión del sistema, el trabajo se enfocó en encontrar solamente oraciones que cumplan con ciertos patrones del lenguaje natural. Dado que los comentarios extraidos de la página corresponden a texto libre, muchas oraciones que contienen información importante se pueden perder ya que la manera en que se encuentran expresadas puede ser muy diversa.

Por ejemplo, no recomendar un lugar se puede expresar de distintas maneras:

- *"Les recomiendo no ir"*
- *"No le recomiendo ir a nadie"*
- *"Yo no lo recomendaria"*
- *"El lugar no es nada recomendable"*

Teniendo un conjunto acotado de reglas quizás no todos los casos mensionados anteriormente puedan ser reconocidos. Estos instancias se verán incluidas en las siguientes refinaciones del sistema.

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

0.2. Descripcion

El desarrollo 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 empezar a analizar los comentarios, fué nescesario definir las palabras relevantes del dominio y clasificarlas, ya que por ejemplo, no cumplen la misma función gramátical *pizza* o *excelente*.

Las palabras se clasificaron en los siguientes grupos:

- Entidad: Define a los sustantivos relevantes del dominio, como ser comidas, atención, ambiente, etc.
- Adjetivo: Son los adjetivos calificativos, estos a su vez se dividieron en:
  - Positivo
  - Neutro
  - Negativo
- Verbo: verbos
- Artículo: articulos
- Palabras vacías: todas aquellas palabras que no están clasificadas en ningun otro grupo.

Dado que el idioma castellano contiene un gran volumen de palabras, no es posible clasificarlas en su totalidad. Con lo que se definió un diccionario acotado de palabras relevantes en nuestro dominio (los comentarios de la guia oleo).

El primer paso para construir el diccionario fue simplemente ordenar por cantidad de apariciones las palabras de nuestro conjunto de datos. Con este procedimiento se logró conseguir una primera aproximación sobre la importancia de las palabras. Con esta lista ordenada se tomaron las primeras 1000 como primer diccionario. Esto se debe a que la clasificación en grupos es de manera manual y lleva mucho tiempo.

Pero dado que el tiempo con que se contaba no era suficiente para caracterizar todas las palabras, se realizó una nueva iteración sobre estas para obtener solamente la raíz para así disminuir el volumen a clasificar y poder obtener un diccionario mas completo. Por ejemplo, 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 segunda un *Adjetivo*. Pero esto fué algo con lo que se aceptó seguir adelante.

Una vez obtenido un diccionario con todas las raices ordenadas por cantidad de apariciones se comenzaron a caracterizar solamente las más importantes. Es necesario aclarar que la clasificación se realizó en 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.

Una vez que se desarrollo el sistema con este diccionario, se decidió enriquecerlo de manera semiautomática. Este proceso será explicado más adelante.

0.2.2. Definición de reglas

Con un diccionario definido, el siguiente paso fué representar los criterios por los cuales que clasificaría a una oración como relevante o no. Esto se realizó mediante el uso de expresiones regulares.

De esta menara se definieron expresiones regulares para la busqueda de las oraciones que cumplen con el criterio mensionado anteriomente. Esto fué realizado manualmente y las expresiones regulares se fueron creando en base al sentido común y pruebas sobre 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 clasificadas y con las reglas, es necesario aplicarlas a los comentarios. Con este fín se desarrolló un sistema informático que realice esta tarea.

El sistema toma como entrada:

- un archivo por cada grupo de palabras,
- un archivo con las reglas definidas y
- los comentarios a evaluar.

El resultado del procesamiento es 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. Por ejemplo:

> *"***La ambientacion es piola aunque** *es* **un lugar bastante ruidoso**. **La pizza sigue siendo rica**. *La desilusion fue el tartufo, que la moza me aseguro que era igual que en Italia y era un simple helado de chocolate. Que lo llamen de otra manera, entonces."*
>
> Resultado:
>
> - R0 un lugar bastante ruidoso - la pizza sigue siendo rica
> - R1 la ambientacion es piola aunque

0.2.4. Recolección de resultados

El siguiente es un ejemplo del resultado completo de sistema aplicado a un comentario:

```
comment: POSITIVO
<1-ops.yaml, user: 17248, service: bueno, food: buena, enviroment: muy bueno, date: 21-05-2010>
Texto orig: La ambientacion es piola aunque es un lugar bastante ruidoso.  La pizza sigue siendo
rica. La desilusion fue el tartufo, que la moza me aseguro que era igual que en Italia y era un
simple helado de chocolate.  Que lo llamen de otra manera, entonces.
result:
                | R0  un lugar bastante ruidoso- la pizza sigue siendo rica-
                | R1  la ambientacion es piola aunque-
```

Donde lo primero que se lista es la clasificación del comentario como positivo/negativo (esta clasificación será explicada con mas detalle mas adelante), las clasificaciones de servicio, comida, ambiente realizadas por el usuario, el comentario original y el resultado de la aplicación de las reglas.

De esta manera se puede observar con claridad el funcionamiento del sistema de extracción de información.

En base a los resultados obtenidos se realizó un análisis preliminar sobre cada comentario, esto fué posible gracias a que los adjetivos fueron divididos en tres tipos

- Positivos,
- Negativos,
- Neutros

De esta manera dado un comentario se hizo un balance entre los adjetivos positivos y negativos que se identificaron como relevantes para dar una idea preliminar del matiz del mismo.

Una vez realizada la categorización de todos los comentarios de un restaurant se creo también un resumen de éste, para tener un una balance general de opiniones.

Para clasificar al restaurant se decidió este acercamiento, en vez de tener en cuenta el porcentaje de adjetivos positivos/negativos en el total de los comentarios, porque lo que se buscó fue encontrar el porcentaje de gente que estuvo conforme con el restaurant, contando a cada persona como un voto positivo o negativo según el matiz general de su comentario.

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

0.2.5. Consideraciones finales

Para empezar, veamos un ejemplo de resultados:

> *"***El ambiente es agradable** *y sencillo,* **la atencion fue muy buena**, *de hecho ayudaron con sugerencias y recomendaciones (quizas por ser dia de semana) .* **La comida fue deliciosa** *y mi acompaniante, que estuvo en Mexico 2 meses, aseguro que* **fue el mejor Chile** *Relleno que probo en Bs As. La pasta de frijoles y el guacamole exquisitos.* **La copa grande** *de* **margarita es muy rica**. *Creo que es el mas fiel a Mexico de todos los restaurantes, ya que tiene el estilo y sencillez de comida y el ambiente que realmente existe en ese pais."*
>
> Resultado:
>
> - R1 el ambiente es agradable - la comida fue deliciosa
> - R2 la atencion fue muy buena - margarita es muy rica
> - R4 la copa grande
> - R5 fue el mejor chile

En este ejemplo se puede observar que lo único que falta identificar es la frase *"La pasta de frijoles y el guacamole exquisitos"* . Esto se debe a que tanto *guacamole* como *frijoles* no están identificados en el diccionario.

Esta falla es un caso típico que se dá cuando se introduce a la base de comentarios uno que haga referencia a platos no identificados previamente. Por este motivo es que se propuso una nueva forma de identificar entidades del dominio de manera más dinámica.

El proceso realizado para esto fue identificar a partir de resultados obtenidos frases relevantes. Luego se llevaron a reglas las estructuras de estas frases, y se procesan los comentarios, esto se verá mejor con un ejemplo:

Se toma la frase *"fue el mejor chile"* , que tiene la siguiente estructura *VERBO ART ADJETIVO ENTIDAD*, si lo que se quiere es reconocer todas las entidades que no se encuentran en el diccionario se remplaza *ENTIDAD* por la palabra clave *XXXX*, de esta manera la regla sería: *VERBO ART ADJETIVO XXXX*.

Al analizar los comentarios del nuevo restaurant con esta regla, el sistema busca todas las oraciones que coincidan de manera tal que *XXXX* sea cualquier palabra no conocida, reportandola como posible *ENTIDAD*.

Existen casos en las que no se desean identificar palabras especificas, por esta razón se incorporó un diccionario de excepciones, las cuales no serán reconocidas con el método anterior.

0.3. Resultados

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

> *''Hacia tiempo que no iba y me agarro nostalgia. Los sabores siguen identicos,* **la cocina es excelente** *y la pizza, a mi entender, de* **la mejor pizza** *italiana que se hace en Buenos Aires, finita, bien a la piedra, y con ingredientes de primera. El carpaccio de* **lomo es buenisimo** *y da para compartir. Creo que es la mejor opcion cuando uno esta por el centro. Comida italiana genuina. Muy recomendable. ''*
>
> Resultado:
>
> - R1 la cocina es excelente- lomo es buenisimo
> - R5 la mejor pizza

Lo primero que se puede apreciar es que las partes identificadas como importantes realmente lo son. Veamos que *"la cocina es excelente"* , *"lomo es buenisimo"* y *"la mejor pizza"* , son tres apreciaciones sobre el restaurant en sí. En cambio oraciones como *"Hacia tiempo que no iba y me agarro nostalgia"* , no lo es, sino que es una comentario personal. Este último tipo de oraciones son las que se desan obviar.

> *''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 este comentario se reconoció casi en su totalidad como relevante. Se puede ver que el dato de que el restaurant queda en la zona de retiro no se identifico, 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 oración relevante: *"comer una rica"* , la regla que se aplicó para identificarla es:

```
RO: (ART)* ENTITY (EMPTY)+ (ADJETIVE)+ (y|ADJETIVE)*;
```

Esto se debe a que la palabra *comer* se reconoció como una entidad y no como un verbo, ya que su raíz es *com*, que también lo es para la palabra *"comida"* , y al ser está ultima una entidad tan importante en el contexto del trabajo se decidio identificarla como tal.

0.4. Conclusiones

Los resultados obtenidos por el sistema implmentados fueron más que satisfactorios. Se puedo apreciar cómo con simples reglas se puede identificar, a apartir de texto libre en lenguaje natural, contenidos relevantes de contenidos fútiles bajo cierto criterio.

Cuando se comnzó con el trabajo de investigación no se tenían grandes expectativas para con el mismo. Pero a medida que el mismo fué creciendo y perfeccionandose se pudo ver cómo se puede crear un sistema funcional que cumpla con el propósito planteado. A medida que se fué iterando tanto sobre las reglas que definen el criterio de relevancia de una frase, como sobre el armado del diccionario de las palabras del dominio se pudo ver como estas refinaciones perfeccionaron el mismo. El progreso del sistema se vió plasmado claramente en los resultados que se fueron viendo.

En las primeras iteraciones a partir de un comentario como:

> *"Para redondear, todo es de mal gusto, salvo la camarera que nos atendio bien, el resto desastre, precios altisimos, la comida no es mas de lo que se puede comer en un bodegon italiano, recomiendo que no vayan, van a salir de mal humor."*

Se obtenían resultados como:

- R0 bodegon italiano , recomiendo
- R3 el resto desastre

Luego, al ir iterando sobre el sistema y a medida que se detectaron las falencias del mismo fué posible que a partir del mismo comentario el sistema detectara correctamente las oraciones relevantes. En el caso planteado anteriormente luego de refinar las reglas y el diccionario fué posible identificar como irrelevante *"bodegon italiano , recomiendo"* , con lo que el resultado obtenido finalmente fué sólo el siguiente:

- R4 el resto desastre

Veamos otro ejemplo; en una de las primeras iteraciones, ante el siguiete comentario:

> *"* **La pizza nos encanta**, *aunque esta vez vino medio fria y bastante quemada. El servicio de mesa es caro (para dos bruschettas de tomate).* **La bebida, carisima** *y los postres, tambien* **son muy caros** *dada la calidad. En fin,* **la pizza es excelente**, *lo demas, deja bastante que desear.* **EL servicio es pesimo**, *y bastante lento: le pedimos a la moza que nos limpiara, por favor,* **la mesa y lo hizo de muy mala gana**. *El mantenimiento de los baños es poco."*

Se obtenían resultados como:

- R0 La bebida , carisima
- R0 son muy caros
- R0 la mesa y lo hizo de muy mala
- R1 La pizza nos encanta
- R1 mesa es caro
- R1 la pizza es excelente
- R1 El servicio es pesimo

Pero luego se consiguió oibtener el siguiente resultado:

> *"* **La pizza nos encanta**, *aunque esta vez vino medio fria y bastante quemada. El servicio de* **mesa es caro** *(para dos bruschettas de tomate). La bebida, carisima y los postres, tambien son muy caros dada la calidad. En fin,* **la pizza es excelente**, *lo demas, deja bastante que desear.* **EL servicio es pesimo**, *y bastante lento: le pedimos a la moza que nos limpiara, por favor, la mesa y lo hizo de muy mala gana. El mantenimiento de los baños es poco."*

Obteniendo resultados más precisos, que es el objetivo que se planteó. Ya que como se explicó en un comienzo se prioriza no identificar información irrelevante como relevante, a riesgo de obviar información que resulte relevante.

Aunque se puede notar que el sistema no es perfecto, sí se puede decir que considerando la totalidad de los casos, el mismo identifica correctamente las partes relevantes en las mayorías de estos.

Fuera de las iteraciones que se realizaron hasta el momento del sistema quedan casos que tiene que ver con la gramática, el contexto o frases verbales. De igual manera resulta interesante ver cómo sólo teniendo en cuenta las oraciones en sí y su configuración se pueden lograr resultados ampliamente satisfactorios, como ser:

" Filo continua sirviendo platos de **la cocina italiana de muy buena calidad***: la pizza estaba* **deliciosa***; el rissoto hecho en el momento, con funghi de primera y arroz arborio, estaba impecable; el semifredo final fue un poema.* **La atencion muy buena***, especialmente en un lugar con tanto movimiento como es Filo,* **La decoracion descontracturada** *y transgresora muy acorde con el sitio. En resumen un restaurante de* **comida italiana de muy buen nivel** *en el centro de BA. Para mi gusto,* **un lugar muy recomendable***, que me da la seguridad de encontrar en el lo que busco cuando quiero comida italiana. "* Resultado:

- R0 la cocina italiana de muy buena calidad
- R0 la pizza estaba deliciosa
- R0 la atencion muy buena
- R0 comida italiana de muy buen nivel
- R0 un lugar muy recomendable
- R4 la decoracion descontracturada

0.5. Trabajo Futuro

A continuación se detallaran algunas ideas que surgieron en el desarrollo del trabajo pero que no se llevaron finalmente a cabo, dejandolas para próximas versiones.

0.5.1. Expresiones Regulares:

Para la evaluación de las expresiones regulares se realiza el siguiente proceso: Se evalúa la oración con todas las reglas, por cada una de estas se recorre las palabras, si estas coinciden con la estructura que se espera se la guarda como resultado y se sigue evaluando la regla con el resto de la frase.

Esto se lleva a cabo con un modulo que implementa la evaluación de expresiones regulares sobre texto, una idea interesante es etiquetar toda la oración con las palabras claves de los distintos grupos y aplicarle una expresión regular. Esto viene implementado en el lenguaje, el problema es que una vez que se obtiene qué parte de la oración de palabras claves cumple con la expresión regular, resulta dificultoso encontrar la parte de la oración original a la que hace referencia. Esta mejora puede hacer que se tengan reglas mas complejas.

0.5.2. Adjetivos:

Para clasificar a los comentarios se cuentan la cantidad de adjetivos positivos y negativos encontrados luego de evaluar las reglas; cada uno de estos tiene peso de un voto, pero esta claro que no es lo mismo decir excelente que horrible y lindo. Por esta razón se puede asignarle un peso diferente a los adjetivos y tener una aproximación mas correcta del matíz del comentario, para la asignación de pesos existe un indicador que se llama *PMI* que indica la "sinominia" de las palabras definida sobre la probabilidad. Esto no fue implementado dado que en el dominio del problema la votación funciona con un alto porcentaje de aciertos.

0.5.3. Interfaz:

Otra idea que no se llegó a implementar se centra en la interfaz del sistema. Dado que la guia oleo consiste una página electrónica hubiera sido interezante acoplar el sistema desarrollado con un parser html que a partir de la página original genere una página igual, pero con las oraciones relevantes remarcadas de algún modo. Así también se podría proveer de un resumen del restaurant en particilar.

Haciendo que la utilización del sistema sea mucho más intuitiva y integrandolo a la página original.