UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

IDENTIFICACIÓN AUTOMÁTICA DE MARCADORES LINGÜÍSTICOS DE INGREDIENTES DE RECETA DE COCINA COSTARRICENSE MEDIANTE MODELOS DE LENGUAJE Y CLASIFICADORES AUTOMÁTICOS

Trabajo final de investigación aplicada sometido a la consideración de la Comisión del Programa de Estudios de Posgrado en Computación e Informática para optar al grado y título de Maestría Profesional en Computación e Informática

SHARON LISETTE CORRALES MONTERO

Ciudad Universitaria Rodrigo Facio, Costa Rica

2019

Dedicatoria

A mis padres y a mi hermano por el apoyo incondicional a lo largo de mi vida y mis estudios. A mis padres, porque su esfuerzo y dedicación me han permitido alcanzar todas mis metas. A mi novio quien ha estado a mi lado apoyandome e impulsandome durante toda la elaboración de este trabajo. A mis abuelos maternos por cuidarme en mi niñez, por impulsarme a mejorar día a día a lograr nuevas metas y por siempre demostrar el orgullo que sentían por cada uno de mis logros a través de sus palabras y de las fotos de cada una de mis graduaciones que adornan su casa.

Agradecimientos

Agradezco a mis padres, a mi hermano y mi novio por todo el amor y el apoyo incondicional brindado durante la elaboración de este trabajo. A mis padres les agradezco especialmente por su apoyo a lo largo de todos mis estudios, por siempre impulsarme a lograr mis metas y por brindarme sus consejos cuando me sentía cansada y frustrada a lo largo de las distintas etapas de mi educación. A mis padres y a mi hermano, les agradezco su presencia continúa durante cada uno de los momentos importantes de mi vida, por su cariño y sus muestras de orgullo ante cada uno de mis logros.

A mi novio le agradezco por su apoyo incondicional desde el comienzo de nuestra relación, por sus comentarios motivacionales para finalizar este trabajo y por estar presente en cada una de las etapas de esta maestría. A mi amiga y compañera de estudios Karen Miranda por acompañarme en cada curso y proyecto de la maestría. Le agradezco por hacer de la maestría una linda experiencia llena de buenos recuerdos y sobre todo por el apoyo mutuo que nos brindamos a lo largo de toda la maestría.

Por último agradezco al Director de mi TFIA, el profesor Edgar Casasola por toda la ayuda y su valiosa guía en la elaboración de este trabajo.

"Este trabajo final de investigacion aplicada fue aceptado por la Comisión del Programa de Estudios de Posgrado en Computación e Informática de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Computación e Informática."

Dr.Jorge Antonio Leoni de León
Representante del Decano
Sistema de Estudios de Posgrado

Dr.Edgar Casasola Murillo
Profesor Guía

M.Sc.Mario Hernández Delgado
Lector

Dra.Gabriela Marín Raventós
Lectora

M.Sc.Marta Eunice Calderón Campos
Representante de la Directora del
Programa de Posgrado en Computación e Informática

Sharon Lisette Corrales Montero
Sustentante

Índice general

Resumen

Un contexto definicional o definitorio es un texto en el cual se encuentran términos con su correspondiente definición. El análisis de estos contextos tiene como enfoque identificar patrones utilizados en el texto para expresar las relaciones conceptuales (vínculos que conectan a los términos entre sí) representativas del dominio en estudio. La identificación de dichos patrones permite extraer la terminología o el léxico específico utilizados en un dominio especializado. Este tipo de análisis es generalmente realizado por expertos lingüistas sobre pequeños conjuntos de datos de manera manual.

Este trabajo representa una continuación de la investigación realizada en [Corrales et al., 2018] con la técnica de reglas y etiquetado de texto. En el presente estudio, se emplean modelos de lenguaje y clasificadores automáticos de texto para identificar marcadores lingüísticos delimitadores de ingredientes de cocina en el dominio de recetas de cocina costarricense. Los clasificadores utilizados fueron *Naive Bayes* y *Support Vector Machine* en la herramienta Weka. Los insumos utilizados para el entrenamiento y prueba de los modelos provenían de los mismos documentos utilizados por [Corrales et al., 2018].

Para cada modelo se evaluó la precisión de los resultados. La mejora en los resultados del problema se evaluó utilizando una prueba de T-student. Al concluir este trabajo, se ofrece, de manera satisfatoria, una mejora en la precisión de los resultados de la identificación de marcadores lingüísticos delimitadores de ingredientes de cocina al usar clasificadores automáticos de texto.

Palabras clave

Modelo de lenguaje, clasificador automático, *Naive Bayes*, *Support Vector Machine*, contextos definicionales

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

Un contexto definicional o definitorio es un texto en el cual se encuentran términos con su correspondiente definición. El análisis de contextos definicionales permite el estudio empírico de la lengua con base en datos reales, provenientes de amplios corpus, relativos a un dominio específico [Soler, 2005]. Sus principales objetivos son sistematizar y estandarizar las formas de definir palabras o términos de un dominio. Además, está fundamentado en la identificación de patrones utilizados en los textos para expresar relaciones conceptuales (es decir, vínculos que conectan a los términos entre sí) representativas de un dominio [Corrales et al., 2018]. Cada patrón funciona como un delimitador o marcador lingüístico al señalar el tipo de relación semántica que existe entre las palabras. Aunque buena parte de la investigación en análisis de contextos definicionales se centra en el afinamiento de los procedimientos, aún no cuenta con metodologías de empleo universal. De hecho, en ocasiones la extración de relaciones definitorias es realizada por expertos de manera manual sobre un conjunto de pequeño de datos previamente clasificados dentro del contexto definicional en estudio. En este trabajo se propone un proceso automático para analizar contextos del dominio de recetas de cocina, concretamente con respecto a la tarea de identificación de ingredientes mediante marcadores lingüísticos.

1.1. Antecedentes

Debido al interés en el análisis de contextos definicionales, surge en el Instituto de Investigaciones Lingüísticas de la Universidad de Costa Rica el proyecto "Análisis de contextos definicionales en corpus de gastronomía tradicional en Costa Rica (CODEGAT)", investigación que tenía como objetivo principal extraer información gastronómica de textos de recetas de cocina costarricense para aportar a la sistematización del conocimiento gastronómico socializado. La selección del dominio de recetas de cocina se debió a la significatividad social de este ámbito de conocimiento. La sistematización propuesta se basaba en la identificación de los ingredientes de la receta, así como en la identificación de los procesos secuenciales y paralelos requeridos para la elaboración del producto. Con la intención de disminuir el tiempo de procesamiento, trabajar

con grandes volúmenes de datos, automatizar la identificación y validación de patrones definicionales, así como la mejora en la recuperación de relaciones conceptuales pertinentes, el equipo de CODEGAT buscó apoyo en estudiantes del Programa de Posgrado en Ciencias de la Computación de la Universidad de Costa Rica. Esta colaboración se realizó como proyecto de laboratorio de los cursos de Recuperación de Información y Procesamiento de Lenguaje Natural. Los resultados fueron presentados como ponencia en el VIII Coloquio Costarricense de Lexicografía de la Universidad de Costa Rica y se plasmaron en el artículo "Análisis de texto para la identificación automática de marcadores lingüísticos definicionales en recetas de gastronomía costarricense" (Apéndice A). Este fue seleccionado para publicación en la revista Kañina, Vol.42 N°3. En relación con el artículo, este contiene la explicación de una técnica semiautomática de análisis consistente en dos fases. La primera fase primeramente se encargada de descargar contenido de la web mediante una araña de búsqueda. Estos textos descargados eran procesados para identificar los verbos y basado en estos verbos se clasificaba el documento como "Receta" o "No Receta" (Figura 1.1).

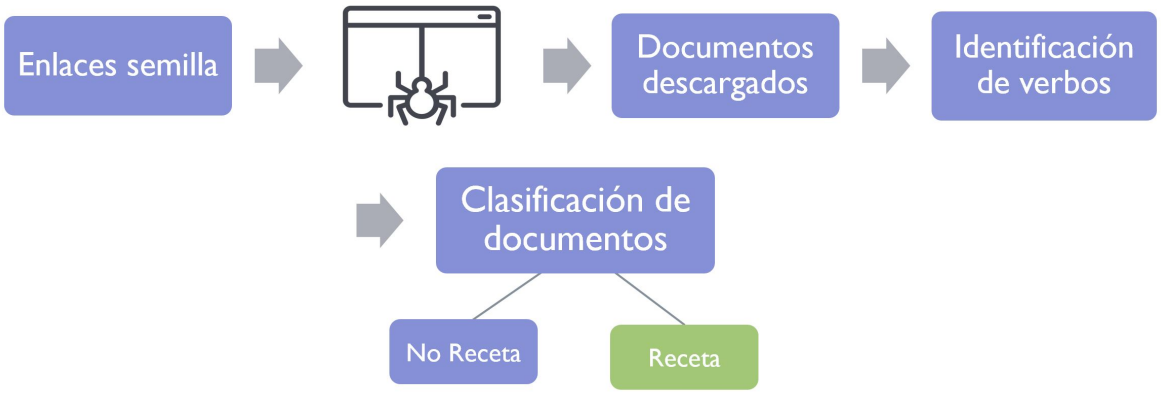


Figura 1.1: Descripción del proceso de la fase 1 de [Corrales et al., 2018]

La segunda se realizó únicamente sobre los documentos clasificados como "Receta" en la fase 1. Esta segunda fase consistió en la identificación de delimitadores lingüísticos (marcadores) para el reconocimiento de los ingredientes de cocina dentro de la receta. En esta fase se realizó la normalización del texto y posteriormente se etiquetó el texto con un etiquetador de partes del discurso (POS Tagger). Seguidamente se utilizaron expresiones regulares apoyadas en las clases semánticas para la identificación de los ingredientes de cocina y se generaron documentos de salida en formato HTML con los marcadores identificados 1.2). El uso de estas expresiones regulares tenía como fin solucionar el problema de las inflexiones nominales y adjetivales, e identificar todas

las diferentes formas de expresar una misma unidad de medida. Sin embargo, El POS Tagger en español etiqueta las unidades de medida en la categoría de nombre común en lugar de numeral de unidad. Por esta razón las expresiones regulares se crearon utilizando la categoría de nombre común para identificar las unidades de medida.

Los resultados obtenidos en la fase 2 evidenciaron que el uso de expresiones regulares apoyadas en etiquetas de partes del discurso no son tan precisas. El uso de la categoría de nombre común en la expresión regular para identificar la unidad de medida, identificaba frases que no contenían unidades de medida. El nombre común para identificar la posición de un ingrediente, provocó la identificación de fragmentos de texto que contenían nombres comunes que no eran ingredientes. Por ejemplo la expresión regular NUM NC AQ (número, nombre común, adjetivo calificativo) identificaba "8 tomates maduros" y "6 bebidas refrescantes".

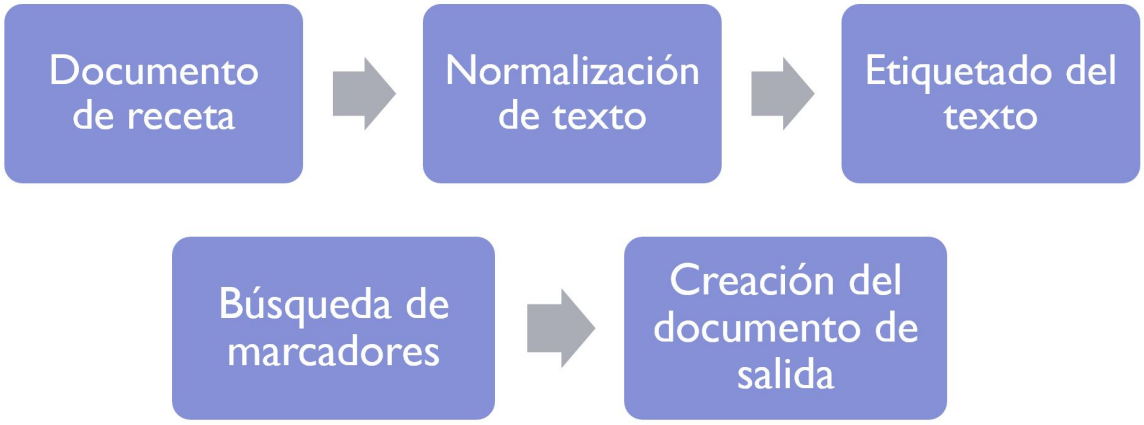


Figura 1.2: Descripción del proceso de la fase 2 de [Corrales et al., 2018]

Ahora bien, la segunda fase de [Corrales et al., 2018] es la base del presente trabajo final de investigación aplicada, en el cual se pretende utilizar técnicas basadas en modelos de lenguaje para la identificación de delimitadores lingüísticos de ingredientes de recetas de gastronomía costarricense. Los modelos de lenguaje son mecanismos probabilísticos que permiten predecir la siguiente palabra en un texto con base en las palabras predecesoras. Después de la creación de los modelos de lenguaje, se realiza una comparación de los resultados de esta técnica con lo propuesto en la segunda fase de la técnica utilizada en [Corrales et al., 2018].

1.2. Organización del presente documento

Este documento se encuentra divido en seis capítulos. Luego del primer capítulo, de carácter introductorio, en el segundo se describen los conceptos teóricos que se utilizaron en el planteamiento del problema de investigación y en la elaboración de la técnica propuesta. En el tercer capítulo se realiza el planteamiento de la pregunta de investigación y se definen los objetivos, alcances y limitaciones. Posteriomente, en el cuarto capítulo se presenta la metodología utilizada en la investigación y se describen en detalle los pasos realizados para obtener el corpus de recetas de cocina costarricense y crear los modelos de clasificación. En el quinto capítulo se brindan los resultados obtenidos para los modelos de clasificación y se comparan los resultados de esta investigación contra los obtenidos en [Corrales et al., 2018]. Finalmente, en el último capítulo se presentan las conclusiones de la investigación y se mencionan los trabajos futuros en los que es posible continuar trabajando a partir de los resultados obtenidos de este análisis.

Capítulo 2

Marco Teórico

En este capítulo se definen detalladamente los conceptos teóricos necesarios para comprender el desarrollo del presente trabajo. Primeramente, se brindan los conceptos de contextos definicionales y marcadores lingüísticos que provienen del área de la lingüística. Posteriormente, se presentan los conceptos relacionados con las áreas de procesamiento de lenguaje natural y recuperacion de información.

2.1. Contextos definicionales

El análisis de contextos definicionales es una línea de investigación lingüística en dominios restringidos cuyo objetivo es clasificar y sistematizar las formas de definir palabras o términos en un área de conocimiento especializado. Posteriormente, esta clasificación y sistematización puede utilizarse en la recuperación de relaciones semánticas entre los términos utilizados en el dominio de interés [Corrales et al., 2018]. Un contexto definicional o definitorio es un texto en el cual se definen términos de un contexto o dominio específico; por ejemplo, este trabajo utilizó textos de recetas de cocina porque las recetas se consideran como definiciones semiespecializadas de caracter procedimental en las que los términos de interés se refieren a los insumos o ingredientes.

2.2. Marcador lingüístico

El término marcador lingüístico se refiere a una palabra, conjunto de palabras o combinación de estructuras gramaticales que frecuentemente funcionan como indicadores de una relación semántica o gramatical entre fragmentos discursivos en un texto. Estos marcadores guían la interpretación del discurso que se transmite en el texto y permiten la identificación de relaciones semánticas de manera rápida y eficaz [Soler, 2005]. Por ejemplo, algunos marcadores lingüísticos identificados en el proyecto CODEGAT para la introducción de los ingredientes de cocina costarricense son:

- 2 barras de
- 3 tazas de

- 2 botellas de
- 3 kilos de
- [X] pequeñas
- 2 [X] batidos
- 1 [X] picada
- 1/2 taza de
- 1 cucharadita de [X][[en]]

donde X corresponde a un conjunto obligatorio y variable de una o más letras, y donde los dobles corchetes cuadrados ([[]]) contienen palabras que pueden o no estar. Por ejemplo, el marcador "1 cucharadita de [X][[en]]", describe a la vez las expresiones "1 cucharadita de orégano" y "1 cucharadita de canela en polvo". Asimismo, otras pautas utilizadas por los expertos en su codificación de reglas se muestran en el cuadro 2.1.

Cuadro 2.1: Simbología consignada en [Corrales et al., 2018] para la codificación inicial

Regla	Significado
[[elemento]]	elemento fijo pero opcional
[X]	elemento variable pero obligatorio
[[x]]	elemento variable y opcional

Por otra parte, la convergencia de un marcador lingüístico y el tipo de información que marca se conoce como patrón lingüístico. Por ejemplo, en este trabajo los patrones lingüísticos corresponden a la convergencia de un marcador lingüístico y un ingrediente de receta de cocina como se muestra en la figura 2.1. En un patrón lingüístico, el marcador lingüístico se puede encontrar al inicio o al final del patrón, o estar divido en el patrón. Cuando, en un patrón lingüístico, el marcador lingüístico se encuentra divido o separado, este patrón se conoce como patrón escindido (Figura 2.2).

2.3. Web *Crawler* o araña de búsqueda

Un web *crawler* es un programa informático para búsqueda e inspección automática de páginas web. Inicialmente, el *crawler* utiliza un conjunto de URL llamados raíces. Por

Figura 2.1: Ejemplo de patrón lingüístico

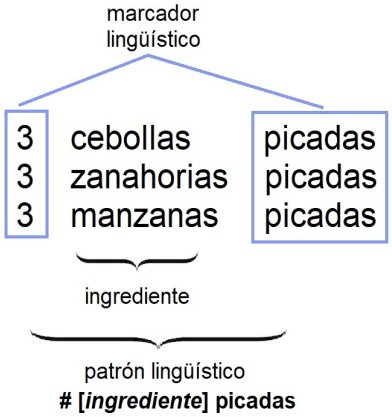


Figura 2.2: Ejemplo de patrón lingüístico escindido

cada una de estas raíces, el *crawler* descarga el contenido de la página web y añade los hiperenlaces contenidos en la página a una lista de direcciones por visitar. El proceso se repite para los enlaces por visitar. La descarga de documentos se configura para finalizar cuando se alcanza un límite de documentos descargados o cuando se alcanza un determinado volumen de información.

2.4. Expresiones regulares

Las expresiones regulares son secuencias de caracteres que representan un patrón de búsqueda. Los caracteres en la secuencia pueden ser literales o metadatos. Los metadatos tienen un significado especial en la expresión regular ya que son utilizados para definir formatos e indicar cuantificación, agrupamiento y alternativas (por ejemplo, picad(o—a—os—as) describe el conjunto de palabras “picado” o “picada” o “picados” o “picadas”) [Fitzgerald, 2012]. La definición de formato se utiliza para verificar que los datos recibidos poseen el formato esperado, por ejemplo para verificar datos de

entrada de correos electrónicos, números de cedúla (# - #### - ####) y fechas (dd/mm/aaaa, dd-mm-aaaa).

2.5. Modelo de lenguaje

Los modelos de lenguaje estadísticos tienen su origen con Claude Shannon [Croft y Lafferty, 2003]. Shannon consideró el lenguaje natural como una fuente estadística, y calculaba qué tan bien se podía predecir la siguiente letra en un texto dado que las n letras anteriores ya se conocían. Además, realizó experimentos con personas para identificar la entropía, es decir, la medida que permite disminuir el nivel de incertidumbre en cómo se comunica la información en el discurso humano del lenguaje inglés. Estos experimentos se basaban en el hecho de que las personas poseen conocimiento estadístico del lenguaje que hablan. Este conocimiento les permite completar oraciones incompletas o truncadas en conversaciones, o completar letras ausentes o mal escritas en pruebas de lectura [Croft y Lafferty, 2003], [Shannon, 1951]. A pesar del surgimiento de modelos de lenguaje con mayor poder de predicción, los estudios de Shannon siguen siendo útiles para tareas de procesamiento de texto. Por otro lado, los modelos de lenguaje siguen siendo utilizados en las áreas computacionales de recuperación de información y procesamiento de lenguaje natural para tareas tales como etiquetado de partes del discurso, reconocimiento del habla, traducción y recuperación de información [Baeza-Yates y Riberito Neto, 2010]. Algunos de estos modelos estadísticos corresponden a bigramas, trigramas y n-gramas.

2.5.1. Bigrama

Los bigramas son grupos de dos palabras y se basan en la probabilidad condicional de que, dada la palabra w_{i-1} la siguiente palabra sea w_i.

$$P(w_i \mid w_{i-1}) = \frac{P(w_{i-1}, w_i)}{P(w_{i-1})}$$

$$= \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

Por ejemplo, la frase “1 cucharadita de sal” posee los bigramas “1 cucharadita”, “cucharadita de” y “de sal”. La probabilidad de que la palabra “1” este seguida por “cucharadita” corresponde a

$$P(cucharadita \mid 1)$$

$$= \frac{P(1, cucharadita)}{P(1)}$$

$$= \frac{count(1, cucharadita)}{count(1)}$$

Del mismo modo, la probabilidad de la frase “1 cucharadita de sal” corresponde a

$$P(1\ cucharadita\ de\ sal)$$

$$= P(cucharadita \mid 1) \cdot P(de \mid cucharadita) \cdot P(sal \mid de)$$

2.5.2. Trigrama

Los trigramas son grupos de tres palabras y, así como los bigramas, se basan en la probabilidad condicional de que, dadas las palabras w_{i-1} y w_{i-2}, la siguiente palabra sea w_i.

$$P(w_i \mid w_{i-1}, w_{i-2}) = \frac{P(w_{i-2}, w_{i-1}, w_i)}{P(w_{i-2}, w_{i-1})}$$

Por ejemplo, la frase “4 onzas de arroz crudo” posee los trigramas “4 onzas de”, “onzas de arroz” y “de arroz crudo”. La probabilidad de que las palabras “onzas de” esten seguidas por “arroz” corresponde a

$$= P(arroz \mid onzas, de)$$

$$= \frac{P(onzas, de, arroz)}{P(onzas, de)}$$

$$= \frac{count(onzas, de, arroz)}{count(onzas, de)}$$

Del mismo modo, la probabilidad de la frase "4 onzas de arroz crudo" corresponde a

$$P(4\ onzas\ de\ arroz\ crudo)$$

$$= P(de \mid 4, onzas) \cdot P(arroz \mid onzas, de) \cdot P(crudo \mid de, arroz)$$

2.5.3. N-grama

Los n-gramas corresponden a grupos de n palabras consecutivas en el texto. Además, utilizan procesos de Markov de orden n para calcular la probabilidad de ocurrencia de una secuencia de palabras S. Por lo tanto, la probabilidad de ocurrencia de un término depende de los *n-1* términos precedentes en el texto [Baeza-Yates y Riberito Neto, 2010].

$$P_n(S) = \prod_{i=1}^{n} P(k_i | k_{i-1}, k_{i-2}, ..., k_{i-(n-1)})$$

Por ejemplo, en la secuencia "2 kilos de posta cocida" la probabilidad de que el texto "2 kilos de posta" esté seguido por el término "cocida" corresponde a

$$= P(2\ kilos\ de\ posta\ cocida)$$

$$= P(2) \cdot P(kilos \mid 2) \cdot P(de|\ kilos, 2) \cdot P(posta \mid de, kilos, 2) \cdot P(cocida|\ posta, de, kilos, 2)$$

2.6. Clasificadores de texto

La definición formal del problema de la clasificación de texto indica que, dada una colección $D = \{d_1, d_2, ..., d_m\}$ de documentos, y un conjunto de clases $C = \{c_1, c_2, ..., c_n\}$, un clasificador asigna una clase a cada documento. Esta clasificación se realiza mediante una función booleana, en la que cada par $[d_i, c_j]$, dados $d_i \epsilon D$ y $c_j \epsilon C$, tiene un valor de 0 o 1. El valor 1 indica que el documento pertenece a la clase y 0 indica no pertenencia [Baeza-Yates y Riberito Neto, 2010]. Ahora bien, los clasificadores de texto pueden ser supervisados, cuando hay entrenamiento, o no supervisados, en caso contrario. Los clasificadores supervisados, a diferencia de los no supervisados, requieren de un conjunto de prueba previamente clasificado por usuarios expertos. Este conjunto de pruebas es utilizado como conjunto de datos de entrada en el proceso de aprendizaje de la función de clasificación. Después del entrenamiento, la función de clasificación es utilizada para clasificar nuevos textos [Baeza-Yates y Riberito Neto, 2010].

2.6.1. Clasificador *Naive Bayes*

Naive Bayes es un clasificador de texto supervisado debido a que requiere ser entrenado previamente. Este clasificador pertenece a la familia de los clasificadores probabilísticos de texto y es utilizado en problemas tales como filtrado de spam y análisis de sentimientos. Por otro lado, está basado en el teorema de *Bayes*. Una de sus presuposiciones es que la ocurrencia de un *feature* es independiente de la ocurrencia de otros *features* [Baeza-Yates y Riberito Neto, 2010]. Con respecto a la representación, este clasificador define el documento $\vec{d_j}$ como un vector de *features* (variables independientes). De esta manera, el clasificador calcula la probabilidad de que un vector de *features* pertenezca a una determinada clase c_k mediante la fórmula

$$P(c_k|\vec{d_j}) = \frac{P(c_k) \times P(\vec{d_j}|c_k)}{P(\vec{d_j})}$$

,

la cual puede, debido a la independencia entre *features*, representarse como

$$\hat{y} = argmax_{k \in \{1,...,K\}} P(c_k) \prod_{i=1}^{n} P(x_i|c_k)$$

donde $\hat{y}$ es la clase más probable para el vector de *features* $\vec{d_j} = (x_1, ..., x_n)$ [Baeza-Yates y Riberito Neto, 2010].

2.6.2. Máquinas de soporte vectorial (SVM)

Las máquinas de soporte vectorial (o SVM, por sus siglas en inglés) son máquinas de aprendizaje supervisado que consideran el problema de aprendizaje como una función no lineal $y = f(x)$ [Wang, 2005]. Con respecto a los datos de entrenamiento, estos corresponden a un conjunto de pares de la forma (x_i, y_i), donde x_i es un vector de *features* y y_i su correspondiente clase o etiqueta [Wang, 2005], [Campell y Ying, 2011]. Ahora bien, dadas dos clases bien separadas, la tarea de aprendizaje busca encontrar un hiperplano tal que los puntos de datos de un lado del plano corresponde a una clase A y los puntos de datos del otro lado del hiperplano corresponden a una clase B (figura 2.3). Los puntos más cercanos a este hiperplano son llamados vectores de soporte o *support vectors*, debido a su importancia en el posicionamiento del hiperplano [Campell y Ying, 2011].

2.7. Archivos ARFF

El software de aprendizaje automático llamada Weka, requiere archivos en formato ARFF (Attribute Relation File Format) para realizar el procesamiento de datos. Este formato de archivos se compone de dos secciones principales. La primera corresponde a la lista de atributos (columnas en el archivo) y su tipo (numérico, nominal, string, fecha o relacional), en la que cada atributo en el conjunto tiene el formato @attribute <nombre-atributo><tipo-dato>(figura 2.4). En el caso del vector de palabras, cada palabra del texto se convierte en un atributo y la lista de atributos es conocida como el diccionario del texto.

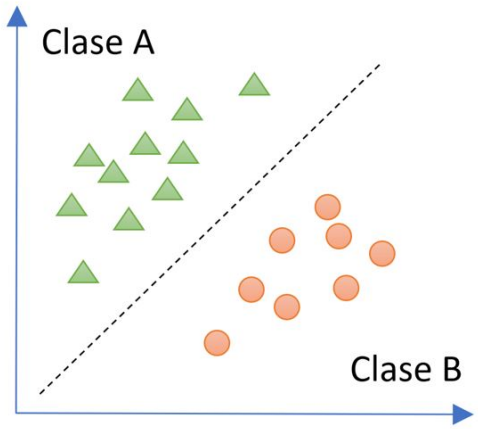


Figura 2.3: Clasificación mediante SVM para dos clases

```
@attribute a1/2 numeric
@attribute aceite numeric
@attribute aceite1 numeric
@attribute aceitu numeric
@attribute aceituna numeric
@attribute aceitunes numeric
```

Figura 2.4: Ejemplo del listado de atributos en archivo ARFF

La segunda sección contiene la declaración de las instancias y su inicio se denota por **@data**. Cada instancia se representa en una línea y contiene los valores de los atributos en la instancia separados por comas, como se muestra en la figura 2.5. Además, el valor enésimo en la instancia corresponde al enésimo atributo declarado en la primera sección del archivo. Para efectos de este trabajo, se utilizaron archivos ARFF escasos en los que cada entrada en la instancia es representada como <indice-atributo><valor-atributo>y los valores 0 no son representados explícitamente. Para el caso de los valores nominales como {Si, No}, el primer valor nominal con índice 0 tampoco se representa explícitamente en la instancia, como se muestra en la figura 2.5.

```
@data
{33 4.627347,117 3.98072,270 0.403393,285 8.123854}
{0 No,7 4.386185,234 3.861174,637 1.748829,890 5.179415}
{0 No,60 2.076482,593 4.486268,833 5.415804}
{0 No,68 4.153562,637 1.748829,662 5.638948}
{0 No,63 4.756558,117 3.98072,270 0.403393,637 1.748829}
{0 No,56 2.993956,90 5.725959,234 3.861174,637 1.748829}
```

Figura 2.5: Ejemplo de la declaración de instancias en archivo ARFF

2.8. Precisión y *recall*

La precisión y el *recall* son medidas utilizadas para evaluar la calidad las instancias recuperadas por un clasificador de texto o sistema de recuperación de información. Estas medidas se calculan en base a la cantidad de instancias recuperadas Re y la cantidad de instancias relevantes Rl [Baeza-Yates y Riberito Neto, 2010].

2.8.1. Precisión

La precisión se refiere a la fracción de instancias recuperadas que son instancias relevantes, es decir, la fracción de verdaderos positivos [Baeza-Yates y Riberito Neto, 2010]. Esta medida es calculada mediante la fórmula

$$Precisión = \frac{Re \cap Rl}{Re}$$

Una alta precisión indica que las instancias recuperadas son en su mayoría instancias relevantes (Figura 2.6). Mientras una baja precisión indica que el clasificador recupera pocas instancias relevantes y las instancias recuperadas son en su mayoría irrelevantes.

2.8.2. *Recall*

El *recall* se refiere a la fracción de instancias relevantes que fueron recuperadas [Baeza-Yates y Riberito Neto, 2010]. Esta medida es calculada mediante la fórmula

$$Recall = \frac{Re \cap Rl}{Rl}$$

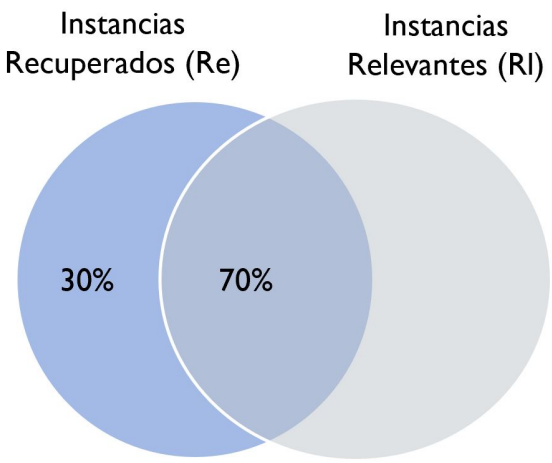


Figura 2.6: Ejemplo de alta precisión

Un alto *recall* indica que la mayor parte de instancias relevantes fueron recuperadas (Figura 2.7). Mientras un bajo *recall* indica que ninguna o pocas instancias relevantes fueron recuperadas.

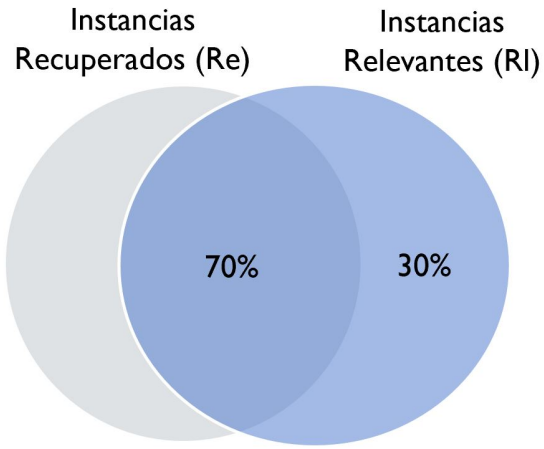


Figura 2.7: Ejemplo de alto *recall*

Al finalizar este capítulo, se posee una conceptualización de los términos utilizados en esta investigación. Ahora que se cuenta con este conocimiento, se procede a describir en el siguiente capítulo el planteamiento de la pregunta de investigación y se definen los objetivos, alcances y limitaciones del presente trabajo.

Capítulo 3

Planteamiento del problema

3.1. Pregunta de investigación

El método propuesto en [Corrales et al., 2018] para identificación de ingredientes de cocina consistió en un sistema basado en expresiones regulares y un etiquetador de las partes del discurso en el texto (verbo, pronombre y adverbio, entre otros). Ahora bien, en el presente trabajo se pretende determinar si el uso de clasificadores automáticos permite mejorar la precisión en la identificación de ingredientes de cocina en comparación con el sistema basado en expresiones regulares, con vistas a tener las mejores opciones para llevar a cabo tal tarea.

3.2. Objetivos

En concordancia con la pregunta de investigación, se pretende determinar si el uso de clasificadores automáticos permite mejorar la precisión en la identificación de ingredientes de cocina en comparación con la técnica de [Corrales et al., 2018]. Para eso, se plantean los siguientes objetivos.

Objetivo General

Evaluar el impacto del uso de clasificadores automáticos sobre la precisión de los marcadores lingüísticos asociados a ingredientes de recetas de cocina en textos.

Objetivos Específicos

1. Generar modelos de clasificación automática utilizando máquinas de soporte vectorial (SVM) para ser aplicados a un corpus de recetas de cocina previamente recolectados de la web, para la identificación de ingredientes.

2. Generar modelos de clasificación automática utilizando clasificación de tipo *Naive Bayes* para ser aplicados a un corpus de recetas de cocina previamente recolectados de la web, para la identificación de ingredientes.

3. Evaluar los resultados obtenidos mediante la aplicación de clasificadores automáticos con base en las técnicas de *Naive Bayes* y máquinas de soporte vectorial (SVM), cada una por separado.

4. Determinar si hay mejora en los resultados de precisión del uso de clasificadores automáticos contra los resultados obtenidos en el proceso semiautomático propuesto en [Corrales et al., 2018] para la identificación de ingredientes de cocina.

3.3. Alcances y limitaciones

Los insumos utilizados para entrenar y generar los modelos fueron los mismos utilizados por [Corrales et al., 2018] en su investigación. Inicialmente, [Corrales et al., 2018] extrajó los documentos mediante una araña de búsqueda de páginas web. Ahora bien, los documentos utilizados en esta investigación corresponden únicamente a los documentos clasificados por [Corrales et al., 2018] en su fase 1 como documentos contenedores de recetas de cocina costarricense. A estos documentos contenedores de recetas, en la fase 2 de [Corrales et al., 2018] se les aplicó un proceso de normalización. Esta normalización consistió de la transformación del texto a minúsculas, eliminación de espacios múltiples entre palabras y transformación de caracteres especiales de fracciones a su forma normal #/#.

Los clasificadores automáticos utilizados fueron *Naive Bayes* y *Support Vector Machine* y se evaluaron utilizando una prueba de T de Student sobre la precisión de los resultados. La metodología utilizada en este trabajo y la evaluación de los resultados obtenidos para los clasificadores automáticos se describen en el siguiente capítulo.

Capítulo 4

Metodología

Tanto el planteamiento de la metodología como su ejecución, así como la exposición que aquí se hace de ella, se organizaron de acuerdo con los pasos necesarios para lograr los objetivos propuestos (figura 4.1). De esta manera, el primer paso corresponde a la recolección de textos pertenecientes al contexto de recetas de cocina costarricense. Esta recolección se realizó mediante el uso de una araña de búsqueda. En el segundo paso se construyeron los conjuntos de prueba y de entrenamiento para los modelos de clasificación. Estos conjuntos tomaron como base los marcadores lingüísticos identificados en [Corrales et al., 2018]. En el tercer paso se procedió a entrenar los modelos de clasificación con el conjunto de datos de entrenamiento. Este entrenamiento se realizó con el uso de la herramienta Weka y los clasificadores automáticos de máquina de soporte vectorial (SVM) y *Naive Bayes*. En el cuarto paso se aplicaron los modelos de clasificación creados sobre el conjunto de pruebas. Por último, en el quinto paso se evaluaron los resultados obtenidos a partir de la clasificación del conjunto de pruebas. En este paso se evaluó la precisión del uso de clasificadores automáticos para la identificación de ingredientes de cocina costarricense y se compararon los resultados de este trabajo con los obtenidos en [Corrales et al., 2018]. En las próximas subsecciones se explica cada paso con mayor detalle.

4.1. Recolección de textos

La recolección de textos de la web se realizó mediante una araña de búsqueda. Esta araña fue configurada con 16 enlaces semilla y se estableció la cantidad de documentos descargados a un límite de 600 documentos. Ahora bien, la selección de estos enlaces semilla se realizó mediante el motor de búsqueda Google. Primeramente, en el motor de búsqueda se ingresó el criterio de búsqueda “recetas de cocina costarricense” para obtener una lista de posibles candidatos a enlace semilla. Posteriormente, se ingresó en cada enlace del conjunto de resultados y se verificó si efectivamente contenían recetas de cocina. En caso de contener la información deseada, el enlace era agregado a la lista de enlaces semilla. Este proceso fue repetido hasta poseer los dieciséis enlaces semilla listados a continuación en el cuadro 4.1.

Cuadro 4.1: Lista de enlaces semilla para el proceso de recolección

No.	Enlace	Fecha acceso
1	http://recetastipicascr.com	Julio 2016
2	http://cocina-consabor.blogspot.com	Julio 2016
3	http://www.southerncostarica.biz/spanish/cat-others/Comidas-Tipicas-de-Costa-Rica/179	Julio 2016
4	https://www.saboresenlinea.com	Julio 2016
5	http://cocina.facilisimo.com/cocina-costarricense	Julio 2016
6	http://www.costaricavipguides.com/guia_recetas_ticas.html	Julio 2016
7	https://cookpad.com/es/buscar/comidas%20de%20costa%20rica	Julio 2016
8	http://www.inforecetas.com/costarica/	Julio 2016
9	http://www.rutas-turisticas.com/recetas_culinarias_cr_costa_rica.html	Julio 2016
10	http://comidastipiguana.blogspot.com/	Julio 2016
11	http://www.dorisgoldgewicht.com/recetas.html	Julio 2016
12	http://www.vivianaentucocina.com/	Julio 2016
13	http://cocinandocontiaflorita.tv	Julio 2016
14	http://www.dcocina.net/?option=com_content&task=category &sectionid=7&id=31&Itemid=34	Julio 2016
15	http://elcucharonylaolla.blogspot.com/2010/07/quien-fuera-tia-florita.html	Julio 2016
16	http://www.labuenacucharacr.com	Julio 2016

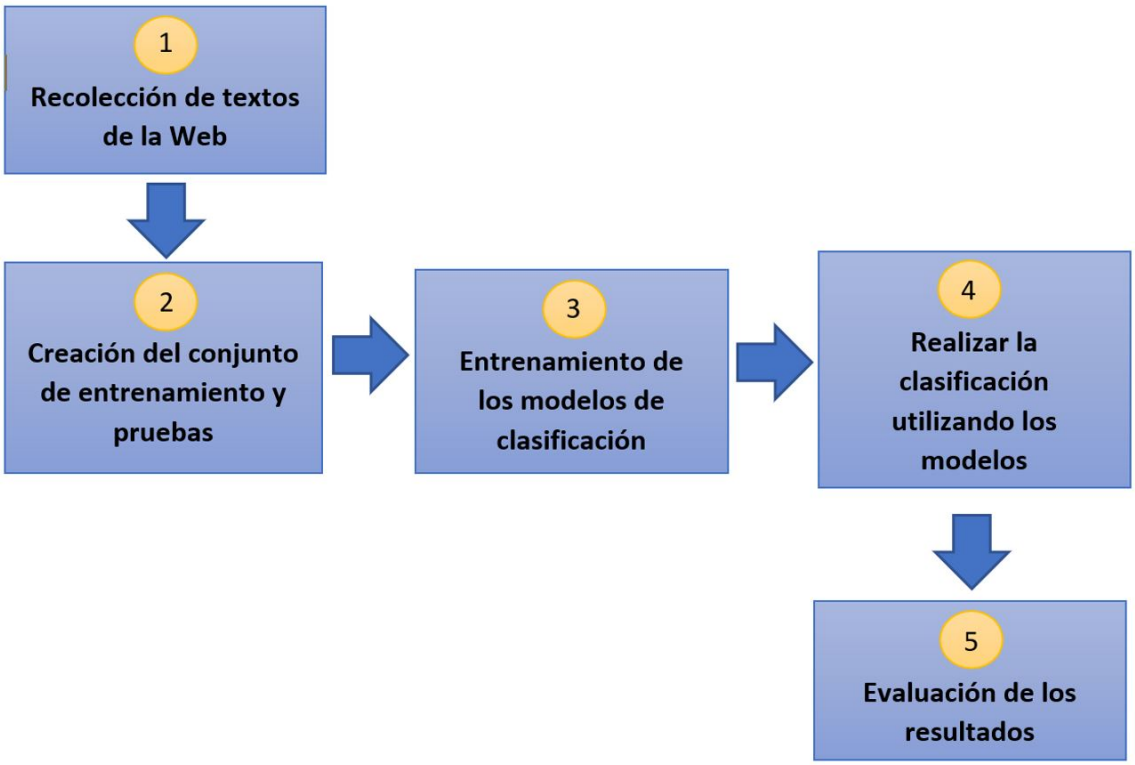


Figura 4.1: Diagrama de la metodología utilizada

4.2. Creación del conjunto de pruebas y entrenamiento

Los conjuntos de prueba y de entrenamiento fueron creados en un proceso semi automático utilizando como archivos fuente los archivos HTML de salida de la aplicación desarrollada por quien escribe el presente trabajo de investigación, tal como se describe en [Corrales et al., 2018]. Esta aplicación tomaba como entrada un grupo de archivos descargados por una araña de búsqueda y los clasificaba en documentos contenedores y no contenedores de recetas de cocina. De los 614 documentos de entrada, solo 383 documentos fueron clasificados como contenedores de recetas de cocina. Posteriormente, en la aplicación mencionada [Corrales et al., 2018] los archivos HTML clasificados como contenedores de recetas se procesaron (normalización del texto, etiquetado de partes del discurso y aplicación de expresiones regulares) para crear archivos HTML de salida en los que se categorizaban los marcadores lingüísticos identificados utilizando etiquetas de HTML (figuras 4.2 y 4.3). Para facilitar la visualización de los marcadores,

estas etiquetas estaban asociadas a un estilo con fuente de color morado como se muestra en la figura 4.4.

Figura 4.2: Procesamiento de archivos en [Corrales et al., 2018]

<hml><head><style>th{font-weight:bold;}table{border-style:none;}span{color:blue}</style><meta charset='UTF-8'></head><body style='font-family:"Courier"'>olla de pozol 13 / 06 / 2016 </br>esta suculenta sopa es la fusiałn de dos recetas ta -picas costarricense : la olla de carne y el pozol . aaideal para compartir en familia! </br>ingrediente </br>**250 gramo de maa-z cascado **- **1 kilo de carne **de res cecina sin grasa cortado en tres trozo - **2 cucharada de salsa lizano **- **2 sobres de sabrosador **de costilla de res criolla - 2 hoja de laurel - **14 taza de agua **- 1 zanahoria grande - 1 cebolla grande - **1 chile dulce rojo **- 3 dientes de ajo - **1 rollo de culantro **- **1 tallo de apio **- **2 taza de agua **- **500 gramo de yuca cortado **en trozo - **500 gramo de camote pelado **y cortado en trozo - **500 gramo de papa pequeaaas **y partido por la mitad - 1 plaatano casi maduro cortado en trozo - **4 taza de caldo **de res - **1 cucharadita de sal **- gota de tabasco </br>

Figura 4.3: Documento HTML resultante en [Corrales et al., 2018]

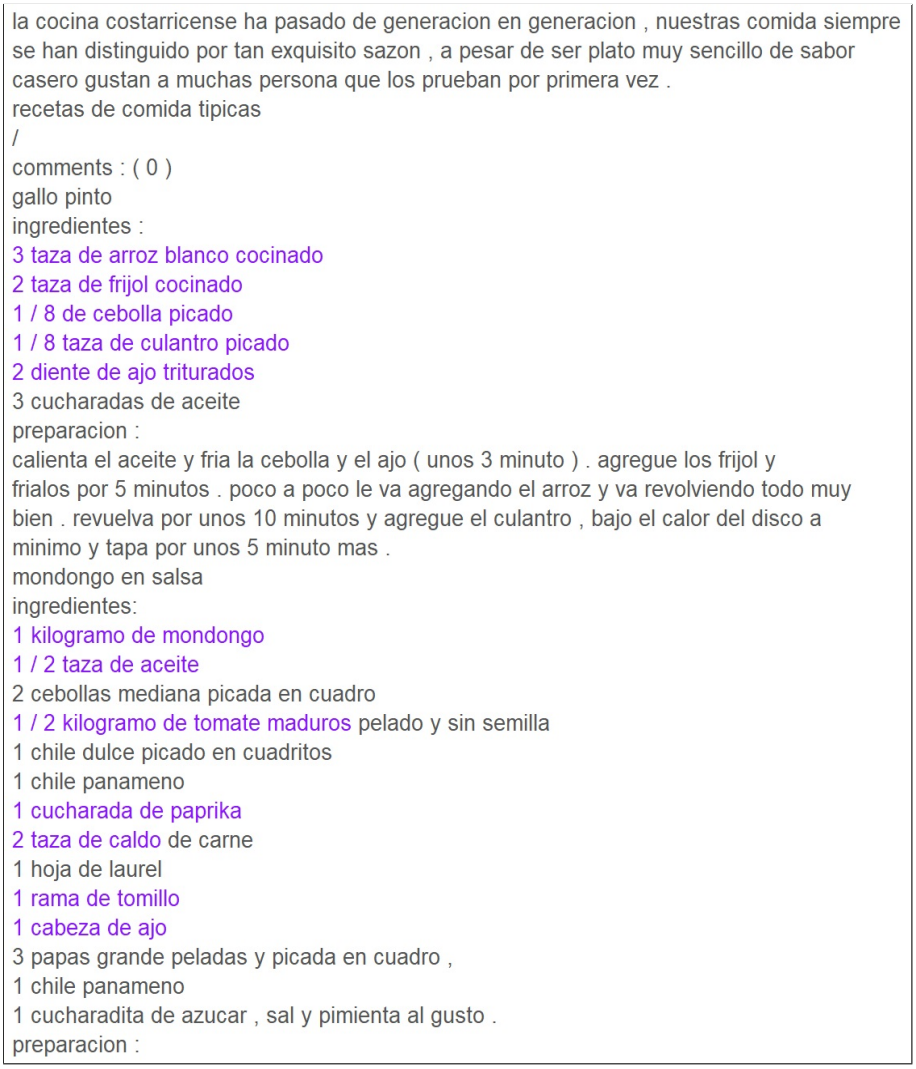


Figura 4.4: Visualización de documento HTML en el navegador.

Ahora bien, el programa propuesto en [Corrales et al., 2018] se modificó para tomar como entradas los archivos HTML con los marcadores identificados y generar nuevas salidas. Estas nuevas salidas eran archivos CSV en el cual cada llínea contenía un marcador lingüístico identificado. Luego, estos archivos fueron unificados y posteriormente se separaron en dos documentos: el 20 % de los 383 documentos de entrada fueron utilizados para el conjunto de entrenamiento y el 80 % restante para el conjunto de pruebas. La selección del 20 % de documentos se realizó por la alta repetición de las mismas unidades de medida en diferentes marcadores lingüísticos. Por esta repetición de datos se requirió de un bajo porcentaje de instancias para lograr generalizar este problema de clasificación.

Para finalizar, los conjuntos fueron etiquetados manualmente con las etiquetas “Si” o

"No", separando esta etiqueta del marcador con el signo de coma. Esta etiqueta indica si el marcador lingüístico cumple o no la función de delimitador de un ingrediente de receta de cocina (Figura 4.5). Al finalizar el proceso de etiquetado, cada archivo contiene una lista de marcadores con su correspondiente etiqueta en la estructura [marcador,clase] (figura 4.6).

Figura 4.5: Proceso de creación del conjunto de entrenamiento y pruebas

marcador, clase
200 drs de arroz,Si
10 recetas con legumbres,No
9 recetas de huevo,No
3 taza de arroz blanco,Si
1/8 taza de cebolla picada,Si
1/2 taza de aceite,Si
6 bebidas refrescantes,No

Figura 4.6: Fragmento de archivo de salida CSV etiquetado manualmente

4.3. Entrenamiento de los modelos de clasificación

Para el entrenamiento de los modelos de clasificación automática Naive Bayes y Máquinas de Soporte Vectorial, se utilizó la plataforma de software de aprendizaje automático Weka con el conjunto de entrenamiento. Antes de ser utilizado para entrenar

los modelos, el conjunto de entrenamiento fue preprocesado tal como se describe en la siguiente subsección (4.3.1).

4.3.1. Preprocesamiento del conjunto de entrenamiento

Para el preprocesamiento del conjunto de entrenamiento, sobre la primera columna correspondiente al marcador lingüístico, se aplicó el filtro NominalToString y posteriomente el filtro de StringToWordVector. El filtro de StringtoWordVector, mediante el proceso de tokenización, convirtió cada marcador en un conjunto de atributos que representan la ocurrencia de la información del texto contenido en las hileras. En segundo lugar, se especificó la segunda columna como la clase de las instancias. Esta columna era un atributo nominal con valores "Si" y "No", donde "Si" indicaba que el marcador lingüístico cumplía la función de delimitador de ingrediente de cocina. Finalizado este preprocesamiento del conjunto de entrenamiento, el resultado se guardó en un archivo de Weka ARFF (figura 4.7).

```
@attribute trucos numeric
@attribute tsp numeric
@attribute turrones numeric
@attribute ufesa numeric
@attribute unicos numeric
@attribute until numeric
@attribute varoma numeric
@attribute velocidad numeric
@attribute vena numeric
@attribute verdura numeric
@attribute video numeric
@attribute w numeric
@attribute yellow numeric
@attribute youtube numeric
@attribute - numeric

@data
{0 No,60 1,789 1,961 1}
{0 No,60 1,790 1,848 1}
{0 No,73 1,270 1,385 1,637 1}
{0 No,60 1,789 1,961 1}
{0 No,60 1,593 1,833 1}
{56 1,453 1,711 1}
{31 1,87 1,270 1,700 1}
{2 1,168 1,270 1,700 1}
{31 1,188 1,577 1}
{31 1,117 1,270 1,364 1,700 1}
```

Figura 4.7: Fragmento del archivo ARFF creado al aplicar filtros sobre el archivo de entrenamiento

4.4. Evaluación de ambos modelos utilizando el conjunto de pruebas

Después de entrenar los modelos de clasificación, se realizó su evaluación utilizando el conjunto de pruebas. Este conjunto contiene 2672 instancias, de las cuales 1585 corresponden a marcadores con función de delimitadores de ingredientes de receta de

cocina. Además, este conjunto fue preprocesado de la misma manera que el conjunto de entrenamiento con los filtros no supervisados de NominalToString y StringToWordVector.

4.5. Evaluación de los resultados

La evaluación de la calidad de resultados se realizó mediante el uso de tablas de confusión, precisión y *recall*. Primeramente se evaluaron los resultados obtenidos por la técnica propuesta en este trabajo, lo cual se hizo comparando entre ellos los resultados logrados con los modelos *Naive Bayes* y SVM. En segundo lugar, se compraron los resultados de la técnica propuesta en este trabajo, basada en modelos de lenguaje, contra los resultados de la técnica propuesta en [Corrales et al., 2018], basada en reglas, etiquetado de texto y expresiones regulares. De esta manera, se determinó cuál de las dos técnicas fue más precisa para resolver el problema de la identificación automática de los ingredientes de cocina en los textos. Finalmente, se realizó una evaluación de la mejora con el uso de modelos de lenguaje y clasificadores automáticos de texto. Los resultados obtenidos y la evaluación de estos se detallan en el siguiente capítulo.

Capítulo 5

Resultados

En este capítulo se presentan los resultados obtenidos en la investigación. En primer lugar se presentan los resultados obtenidos en el entrenamiento de los modelos de clasificación. En segundo lugar se detallan los resultados obtenidos en evaluación de los modelos de clasificación. Seguidamente se presentan la comparación de resultados entre el modelo de SVM y *Naive Bayes* y la comparación de resultados de los modelos de clasificación y la técnica de [Corrales et al., 2018]. Por último se presenta la evaluación de la mejora en los resultados de precisión al utilizar modelos de clasificación automáticos.

5.1. Entrenamiento de los modelos de clasificación

En esta sección se muestran los resultados obtenidos en el entrenamiento de cada modelo de clasificación. Este entrenamiento se realizó utilizando el conjunto de entrenamiento, el cual consistía de 702 instancias. Los resultados se muestran en términos de precisión y *recall* para cada una de las clases.

5.1.1. Entrenamiento del modelo *Naive Bayes*

El entrenamiento del clasificador *Naive Bayes* con el conjunto de entrenamiento generó un modelo con un porcentaje del 93.87 % de instancias clasificadas correctamente y un 6.13 % de instancias clasificadas erróneamente. Es decir, de las 702 instancias con las cuales se entrenó el modelo, solo 43 instancias fueron clasificadas erróneamente, como se puede visualizar en el cuadro 5.1. Esta clasificación errónea corresponde a 9 falsos positivos y 34 falsos negativos.

Cuadro 5.1: Tabla de confusión del entrenamiento del modelo *Naive Bayes*

	Clasificado como	
	Si	No
Si	495	9
No	34	164

Por otro lado en el cuadro 5.2 se muestra un resumen de las estadísticas del modelo

en cuanto a valor de concordancia y media absoluta del error. El valor de concordancia o estadístico kappa de 0.84 nos indica una concordancia cercana a perfecta. En el cuadro 5.3 se puede visualizar que el modelo obtuvo un 0.94 de precisión y 1 de *recall* para la clase "Si" y un 0.95 de precisión y 0.98 de *recall* para la clase "No", lo cual indica que se recuperaron casi todas las instancias relevantes.

Cuadro 5.2: Resumen estadístico del modelo *Naive Bayes*

Estadístico	Valor
Kappa	0.84
Media absoluta del error	0.08
Raíz cuadrada de la media absoluta del error	0.25

Cuadro 5.3: Precisión y *Recall* por clase del modelo *Naive Bayes*

	Precisión	*Recall*
Si	0.94	1
No	0.95	0.98

5.1.2. Entrenamiento del modelo de SVM

Para el entrenamiento del clasificador de SVM, se seleccionó en la plataforma de Weka la función Sequential Minimal Optimization (SMO). El SMO es una variante de entrenamiento del problema de SVM que particiona el conjunto de entrenamiento en conjuntos más pequeños. Ahora bien, el modelo generado con este clasificador ofrece un porcentaje del 99.15 % de instancias clasificadas correctamente y un 0.85 % de instancias clasificadas erróneamente. Es decir, de las 702 instancias con las cuales se entrenó el modelo, solo 6 instancias fueron clasificadas erróneamente, como se puede visualizar en la tabla de confusión (cuadro 5.4). Esta clasificación errónea corresponde a 1 instancia de falso positivo y 5 instancias de falso negativo.

En cuanto a estadísticas, en el cuadro 5.5 se desglosan las estadísticas generales del modelo, y en el cuadro 5.6 se desglosan la precisión y *recall* del modelo por clase. Como se puede observar, el valor de concordancia o estadístico kappa corresponde a un 0.98 (Cuadro 5.5). Este valor, al ser mayor que 0 y cercano a 1, nos indica que la concordancia es cercana a perfecta. Por otro lado, la precisión de ambas clases corresponde a 0.99

Cuadro 5.4: Tabla de confusión del entrenamiento del modelo de SVM

	Clasificado como	
	Si	No
Si	503	1
No	5	193

y 1, respectivamente, lo cual indica que la mayoría de las instancias recuperadas eran relevantes. Por otro lado, los valores de *recall*, de 1 y 0.98, indican que se recuperaron casi todas las instancias relevantes, tal como se muestra en el cuadro 5.6.

Cuadro 5.5: Resumen estadístico del modelo de SVM

Estadístico	Valor
Kappa	0.98
Media absoluta del error	0.01
Raíz cuadrada de la media absoluta del error	0.09

Cuadro 5.6: Precisión y *Recall* por clase del modelo de SVM

	Precisión	*Recall*
Si	0.99	1
No	1	0.98

5.2. Evaluación de ambos modelos utilizando el conjunto de pruebas

En esta sección, se presentan los resultados obtenidos al evaluar los modelos. Esta evaluación se realizó utilizando el conjunto de pruebas, el cual fue preprocesado de la misma manera que el conjunto de entrenamiento. Este conjunto de pruebas consiste de

5.2.1. Evaluación del modelo *Naive Bayes*

La evaluación del modelo Naive Bayes con el conjunto de pruebas mostró un 92 % de instancias clasificadas correctamente. Por lo tanto, el porcentaje de error fue de 8 %.

En este porcentaje de error están representados un 3.79 % de error en la clasificación de marcadores de ingredientes de cocina y un 14.17 % de error para los marcadores que no cumplen como delimitadores de ingredientes de cocina (cuadro 5.7).

Cuadro 5.7: Tabla de confusión de las pruebas del modelo *Naive Bayes*

	Clasificado como	
	Si	No
Si	1525	60
No	154	933

En cuanto a estadísticas, en los cuadros 5.8 y 5.9 se desglosan las estadísticas generales del modelo, la precisión y el *recall* del modelo por clase. Basados en los valores de la precisión y el *recall*, la clasificación de los marcadores que son delimitadores de ingredientes de receta de cocina fue correcta en más ocasiones que para aquellos marcadores correspondientes a la clase "No".

Cuadro 5.8: Resumen estadístico de las pruebas del modelo *Naive Bayes*

Estadístico	Valor
Kappa	0.82
Media absoluta del error	0.09
Raíz cuadrada de la media absoluta del error	0.27

Cuadro 5.9: Precisión y *Recall* por clase de las pruebas del modelo *Naive Bayes*

	Precisión	*Recall*
Si	0.90	0.96
No	0.94	0.86

5.2.2. Evaluación del modelo de SVM

La evaluación del modelo de *Support Vector Machine* con el conjunto de pruebas mostró un 93.38 % de instancias clasificadas correctamente y un error del 6.62 %. Este porcentaje de error esta representado como un 2.97 % de error en la clasificación de marcadores de ingredientes de cocina y un 11.96 % de error para los marcadores que no

cumplen como delimitadores de ingredientes de cocina, como se muestra en el cuadro 5.10.

Cuadro 5.10: Tabla de confusión de las pruebas del modelo de SVM

	Clasificado como	
	Si	No
Si	1538	47
No	130	957

En cuanto a estadísticas, en los cuadros 5.11 y 5.12 se desglosan las estadísticas generales del modelo, la precisión y el *recall* del modelo por clase. Basados en los valores de la precisión y el *recall*, los marcadores que son delimitadores de ingredientes de receta de cocina son clasificados de manera correcta en más ocasiones que aquellos marcadores correspondientes a la clase "No". En otras palabras, para la clase "Si" se recuperó el 97 % de las instancias relevantes.

Cuadro 5.11: Resumen estadístico de las pruebas del modelo de SVM

Estadístico	Valor
Kappa	0.86
Media absoluta del error	0.07
Raíz cuadrada de la media absoluta del error	0.26

Cuadro 5.12: Precisión y *Recall* por clase de las pruebas del modelo de SVM

	Precisión	*Recall*
Si	0.92	0.97
No	0.95	0.88

5.3. Comparación de los resultados

En esta sección, se presentan las comparaciones entre modelos en términos de los resultados de precisión y *recall*. En la primera parte, se realizó una comparación de los resultados del modelo de *Naive Bayes* con los del modelo de SVM. En la segunda parte, se realizó la comparación de los resultados de la técnica propuesta en [Corrales et al., 2018] con los resultados de los modelos de clasificación utilizados en este trabajo.

5.3.1. Comparación de los resultados entre los modelos *Naive Bayes* y SVM

Según los resultados obtenidos en la validación de ambos modelos, el modelo de SVM obtuvó mejores resultados que el modelo de *Naive Bayes* para la identificación de marcadores como delimitadores de ingredientes de cocina. Esto debido a que, tanto la precisión como el *recall* del modelo de SVM presentan mejores resultados en la clasificación de ambas clases, como se muestra en el cuadro 5.13. Por otra parte, podemos observar que ambos modelos presentan mayor precisión en los resultados de la clase "No" que en los resultados de la clase "Si". Mientras para el *recall*, en ambos modelos los resultados son mejores para la calase "Si".

Cuadro 5.13: Tabla de confusión de las pruebas del modelo de SVM

	Precisión		*Recall*	
	Bayes	SVM	Bayes	SVM
Si	0.90	0.92	0.96	0.97
No	0.94	0.95	0.86	0.88

Para evaluar los modelos, se utilizó una prueba T de Student con un nivel de significancia del 5 %, para la hipótesis nula de que el uso del modelo de SVM presenta una mejora en la precisión de los resultados en contraste con el modelo de *Naive Bayes*. En cuanto a las muestras, se eligieron muestras independientes con varianzas distintas y con *n-1* grados de libertad para ambas muestras. En otras palabras, se tomaron 12 documentos preprocesados pertenecientes al conjunto de pruebas. De estos documentos se eligieron seis documentos al azar para ser procesados con el modelo de clasificación de SVM y otros seis documentos diferentes para ser procesados con el modelo de *Naive Bayes*. Los resultados de este procesamiento en términos de precisión se muestran en el cuadro 5.14.

Para el modelo de SVM se tenía una media $\mu_1 = 0{,}96$ y una varianza $S_1^2 = 0{,}003$, y para el modelo de *Naive Bayes* se tenía una media $\mu_2 = 0{,}92$ y una varianza $S_2^2 = 0{,}007$. En la siguiente fórmula calculamos el estadístico de prueba t con $n_1 + n_2 - 2$ grados de libertad.

$$t_0 = \frac{\mu_1 - \mu_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{0{,}96 - 0{,}92}{\sqrt{\frac{0{,}003}{6} + \frac{0{,}007}{6}}} = 0{,}98 \tag{5.1}$$

Cuadro 5.14: Precisión de las muestras tomadas para SVM y Bayes

No. Documento	Precisión SVM
147	0.88
160	0.95
390	0.90
459	1
271	1
572	1
Media	0.96

No. Documento	Precisión Bayes
142	0.96
456	1
204	0.88
444	1
263	0.79
371	0.91
Media	0.92

Para una confianza del 95 % con 10 grados de libertad, el valor crítico corresponde a $t_{\alpha/2} = t_{0,025} = 2{,}2281$. Dado que $t_{\alpha/2} > t_0$ (5.1), se rechazó la hipótesis nula. Por lo tanto, podemos concluir que el uso del modelo de SVM no presenta mejora en la precisión de los resultados en contraste con el modelo de *Naive Bayes*.

5.3.2. Comparación de los resultados de los modelos *Naive Bayes* y SVM con la técnica propuesta en [Corrales et al., 2018]

La técnica propuesta en [Corrales et al., 2018] obtuvo un 53 % de instancias identificadas correctamente, mientras que la técnica propuesta en el presente trabajo obtuvo un 92 % de instancias identificadas correctamente para el modelo *Naive Bayes* y de 93.40 % para el modelo de SVM. En cuanto al *recall*, el uso de clasificadores automáticos presenta casi el doble del valor que la técnica de [Corrales et al., 2018]. Esto indica que la mayoría de los marcadores identificados en [Corrales et al., 2018] son falsos positivos o marcadores cuya función no es delimitar ingredientes de receta de cocina (52 %), mientras este valor en el modelo SVM tan solo es del 3 %, tal como se muestra en el cuadro 5.15.

Igualmente, en el cuadro 5.15 se puede observar una mejor precisión para los modelos de lenguaje (0.90 para *Naive Bayes* y 0.92 para SVM) en comparación con el modelo de [Corrales et al., 2018] (0.70). Esto indica que aproximadamente el 10 % de los marcadores identificados como delimitadores de ingredientes de receta de cocina con los modelos de lenguaje fueron identificados incorrectamente, mientras que este mismo valor en [Corrales et al., 2018] es de un 30 %,es decir, tres veces más que los modelos de lenguaje.

Cuadro 5.15: Comparación de la técnica de [Corrales et al., 2018] y la técnica propuesta en este trabajo

Técnica	Modelo	Precisión	*Recall*
[Corrales et al., 2018]	N/A	0.70	0.48
Modelos de lenguaje y clasificadores automáticos	Bayes	0.90	0.96
	SVM	0.92	0.97

5.4. Evaluación de la mejora

En el cuadro 5.15, se compararon los resultados de precisión y *recall* obtenidos por los modelos de clasificación y la técnica de [Corrales et al., 2018] sobre el conjunto de pruebas. De estos modelos, el modelo de SVM presentó mejores resultados que el modelo de *Naive Bayes*. Por lo tanto, se eligió el modelo de SVM para realizar la evaluación de la mejora en los resultados de precisión al usar clasificadores automáticos para la identificación de ingredientes de cocina.

La mejora se evaluó utilizando una prueba T de Student con un nivel de significancia del 5 % para la hipótesis nula de que el uso de modelos de clasificación presenta una mejora en la precisión de los resultados en contraste con la técnica de [Corrales et al., 2018]. En cuanto a las muestras, se eligieron muestras independientes con varianzas distintas y con *n-1* grados de libertad para ambas muestras. Los resultados de este procesamiento en términos de precisión se muestran en el cuadro 5.16.

Cuadro 5.16: Precisión de las muestras tomadas

No. Documento	Precisión SVM
147	0.88
160	0.95
390	0.90
459	1
271	1
572	1
Media	0.96

No. Documento	Precisión [Corrales et al., 2018]
592	0.67
144	0.89
614	0.63
190	0.79
183	0.63
267	0
Media	0.60

Para el modelo de SVM se tenía una media $\mu_1 = 0{,}96$ y una varianza $S_1^2 = 0{,}003$, y para la técnica de [Corrales et al., 2018] se tenía una media $\mu_2 = 0{,}60$ y una varianza $S_2^2 = 0{,}097$. En la siguiente fórmula calculamos el estadístico de prueba t con $n_1 + n_2 - 2$ grados de libertad.

$$t_0 = \frac{\mu_1 - \mu_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{0{,}96 - 0{,}60}{\sqrt{\frac{0{,}003}{6} + \frac{0{,}097}{6}}} = 2{,}79 \tag{5.2}$$

Para una confianza del 95 % con 10 grados de libertad, el valor crítico corresponde a $t_{\alpha/2} = t_{0{,}025} = 2{,}2281$. Dado que $t_{\alpha/2} < t_0$ (5.2), se aceptó la hipótesis nula. Por lo tanto, podemos concluir que el uso de modelos de clasificación brindan una mejora con respectos a la técnica de [Corrales et al., 2018] en la precisión para la identificación de ingredientes de cocina.

5.5. Análisis de casos particulares

En esta sección se presenta el análisis de dos casos particulares clasificados de manera incorrecta por el modelo de SVM. Primeramente, para la clasificación errónea en la clase "No", analizaremos el texto "6 palito de cangrejo". Posteriomente, para la clasificación incorrecta en la clase "Si", analizaremos el texto "8 plato de relleno".

6 palito de cangrejo

En este texto, las palabras "palito" y "cangrejo" no se encontraban en los marcadores lingüísticos del conjunto de entrenamiento. Además "6" se encontraba principalmente en textos clasificados como no delimitadores de ingredientes de receta de cocina. Por lo tanto, la presencia de "6" y el peso cero de las otras palabras afectó en la clasificación del texto y produjó un error de predicción del 0.91.

8 plato de relleno

En este caso, "relleno" no se encontraba en el conjunto de entrenamiento y "8" se encontraba en ambos clases en proporciones similares. No obstante, "de" se encontraba principalmente en marcadores lingüísticos de ingredientes de cocina. Debido a estas condiciones, el texto fue erróneamente clasificado como un delimitador.

Las conclusiones del presente trabajo y los posibles trabajos futuro basados en los resultados obtenidos en este trabajo se presentan en el siguiente capítulo.

Capítulo 6

Conclusiones y trabajo futuro

En el presente trabajo se propuso evaluar el impacto del uso de clasificadores automáticos y modelos de lenguaje sobre la precisión en la identificación de marcadores lingüísticos como delimitadores de ingredientes de cocina. Este objetivo se logró mediante una serie de tareas descritas en la metodología. En concreto, en el capítulo 5, particularmente en el apartado 5.4 "Evaluación de la mejora", se compararon los resultados de [Corrales et al., 2018] con la técnica del presente trabajo, basada en modelos del lenguaje y clasificadores automáticos de texto. Esta comparación logró evidenciar un impacto positivo de la técnica propuesta en esta investigación para la identificación de los marcadores en relación con la técnica de [Corrales et al., 2018]. En el siguiente apartado se detalla, en orden de planteamiento, cómo se lograron cada uno de los objetivos específicos de este trabajo. En el segundo apartado se indican posibles trabajos futuros que podrían desarrollarse a partir de estos resultados.

6.1. Conclusiones

En este trabajo se generaron conjuntos de datos de marcadores lingüísticos, etiquetados con las clases "No" o "Si" para indicar si el marcador lingüístico cumplía función de delimitador de ingrediente de cocina. Además, se logró generar modelos de clasificación con *Naive Bayes* y máquinas de soporte vectorial con precisiones del 90 % y 92 % respectivamente. Al evaluar los resultados de estos modelos contra los resultados de la técnica de [Corrales et al., 2018], el uso de clasificadores de texto presentó una mejora en la precisión de los resultados del 22 %.

Adicionalmente, se evidenció un aumento del 49 % en el *recall* para el mismo problema al usar los modelos de lenguaje. Esto indica que la técnica propuesta en este trabajo identificó de manera correcta más marcadores lingüísticos y se produjo una disminución en la cantidad de textos identificados erróneamente como marcadores lingüísticos.

6.2. Trabajo futuro

Dada la prueba de T student realizada para comparar la precisión de los resultados entre los clasificadores de *Naives Bayes* y SVM, el uso de un clasificador no presenta mejora en la precisión de los resultados en comparación con el otro clasificador. Además, en esta investigación el uso de clasificadores simples presentó una alta precisión para el problema en estudio. Por lo tanto, no se considera necesario utilizar clasificadores más complejos y que requieran de mayor tiempo para ser entrenados porque la mejora en la precisión no va a representar una diferencia significativa en comparación con los resultados obtenidos en esta investigación. Sin embargo, podría evaluarse el efecto que tiene el uso de diferentes métodos de asignación de pesos a los términos, sobre la calidad de los resultados.

Bibliografía

[Baeza-Yates y Riberito Neto, 2010] Baeza-Yates, R. y Riberito Neto, B. (2010). *Modern Information Retrieval: The Concepts and Technology Behind Search*. Addison-Wesley Publishing Company, USA, 2nd edition.

[Campell y Ying, 2011] Campell, C. y Ying, Y. (2011). *Learning with Support Vector Machines*. Morgan & Claypool Publishers.

[Corrales et al., 2018] Corrales, S., Miranda, K., Casasola, E., Leoni, A., y Hernández, M. (2018). Análisis de texto para la identificación automática de marcadores lingüísticos definicionales de recetas de gastronomía de costa rica. *Revista Káñina, Universidad de Costa Rica*, 42(3).

[Croft y Lafferty, 2003] Croft, B. y Lafferty, J. (2003). *Language Modeling for Information Retrieval*. Springer Netherlands.

[Fitzgerald, 2012] Fitzgerald, M. (2012). *Introducing Regular Expressions*. O'Reilly Media Inc.

[Shannon, 1951] Shannon, C. (1951). Prediction and entropy of printed english. *Bell System Technical Journal*.

[Soler, 2005] Soler, V. (2005). Patrones linguísticos para la búsqueda de información conceptual en el corpus textual especializado de la cerámica txtcera.

[Wang, 2005] Wang, L. (2005). *Support Vector Machines: Theory and Applications*. Springer-Verlag Berlin Heidelberg.

A P É N D I C E S

Apéndice A

Artículo “Análisis de texto para la identificación automática de marcadores lingüísticos definicionales en recetas de gastronomía de Costa Rica”

ANÁLISIS DE TEXTO PARA LA IDENTIFICACIÓN AUTOMÁTICA DE MARCADORES LINGÜÍSTICOS DEFINICIONALES EN RECETAS DE GASTRONOMÍA DE COSTA RICA

Text analysis for automatic identification of definitional linguistic markers in Costa Rican gastronomy recipes

Sharon Corrales[*], *Karen Miranda*[**], *Édgar Casasola*[***], *Antonio Leoni*[****], *Mario Hernández*[*****]

RESUMEN

El análisis de contextos definicionales permite clasificar y sistematizar las informaciones definicionales pertenecientes a un dominio específico y, posteriormente, identificar estándares de las formas en que se definen las palabras y términos en tal dominio. En este artículo se describe el proceso realizado para automatizar el análisis de contextos definicionales en el dominio gastronómico de Costa Rica. La labor se realizó mediante el uso de herramientas computacionales para el procesamiento de lenguaje natural. La automatización permite el análisis sobre grandes volúmenes de datos y obtener resultados en menos tiempo del requerido por el análisis manual. Ahora bien, el procedimiento consta de dos módulos, uno de clasificación de documentos en textos con recetas o sin ellas, y un segundo módulo de identificación de los ingredientes de cocina con base en patrones lingüísticos formales.
Palabras clave: análisis lingüístico de recetas, análisis de contextos definicionales, patrones definicionales, marcadores definicionales, procesamiento del lenguaje natural.

ABSTRACT

The analysis of definitional contexts allows to classify and systematize the definitional information belonging to a specific domain, and then to identify standards for the forms in which words and terms are defined in this domain. This paper describes the process implemented to automate the analysis of definitional contexts in the gastronomy domain in Costa Rica. The automation was done by using computational tools for natural language processing. The automation enables analysis of large quantities of data and results in less time than required by manual analysis. Automation consists of two modules, the first one is for the classification of documents in texts with or without recipes and the second one is for the identification of recipe ingredients based on formal linguistic patterns.
Key words: linguistic analysis of recipes, analysis of definitional contexts, definitional patterns, definitional markers, natural language processing.

[*] Estudiante de la Maestría en Computación, UCR. Correo e.: sharoncm.1691@gmail.com.

[**] Estudiante de la Maestría en Computación, UCR. Correo e.: karenmh09@gmail.com.

[***] Escuela de Computación e Informática y Posgrado en Computación, UCR. Correo e.: casasola@gmail.com.

[****] Escuela de Filología, Lingüística y Literatura y Posgrado en Lingüística, UCR. Correo e.: a.leoni@me.com.

[*****] Programa Estudios de Lexicografía, UCR. Correo electrónico: pdfmario@gmail.com.

1. Introducción

El análisis de contextos definicionales o definitorios (en adelante, análisis de CD) es una línea de investigación que permite, en primer lugar, clasificar y sistematizar las informaciones definicionales relativas a un dominio restringido. Posteriormente, esa organización conceptual puede servir tanto para la recuperación de relaciones semánticas definitorias a partir de textos como para la estandarización de las formulaciones definicionales del dominio de especialidad estudiado (cf. Alarcón 2003; Alcina y Valero 2008; Sierra, Alarcón y Aguilar 2006).

A causa del interés en las posibilidades de este tipo de estudios, surge el proyecto "Análisis de contextos definicionales en corpus de gastronomía tradicional en Costa Rica (CODEGAT)", investigación que pretende examinar la información gastronómica presente en textos de recetas costarricenses con el fin último de aportar a la sistematización del conocimiento gastronómico socializado.

Parte importante de esa sistematización es la adecuada identificación de la lista de productos/ingredientes que serán objeto de las diversas acciones y procesos, así como la precisa descripción de las tareas paralelas y secuenciales en las que aquellos se utilizarán. Desde el enfoque del análisis de CD, que es el que aquí seguimos, lo fundamental es identificar las formas recurrentes que se utilizan efectivamente en los textos para la expresión de las relaciones conceptuales pertinentes (cf. Sierra 2009, Valero 2009, Valero y Alcina 2009, Soler 2005, Sierra y Alarcón 2002). A esas formas recurrentes se les llama, en esta perspectiva investigativa, "patrones definicionales", cada uno de los cuales asocia una **clase de contenidos semánticos** con una **clase de formas que sirven para introducirlos** dentro de la cadena textual (y que funcionan como marcadores, señalizadores, indicadores).

A pesar de tener ya varios lustros en desarrollo, la línea de análisis de CD no cuenta aún con paradigmas metodológicos de empleo universal[1]. Sin embargo, una característica esencial de su planteamiento es la automatización de los procedimientos de identificación y validación de los patrones definitorios propuestos, así como de la recuperación de las relaciones conceptuales pertinentes. Esta automatización permite trabajar sobre grandes volúmenes de datos y obtener resultados en menos tiempo que el requerido por el análisis manual. Debido a lo anterior, el equipo de trabajo de CODEGAT incluye tanto a lingüistas como a especialistas con conocimientos en procesamiento del lenguaje natural.

Ahora bien, en relación con el proceso del análisis de texto, este se dividió en dos módulos (v. figura 1). El primer módulo corresponde a la clasificación de los documentos en aquellos que contienen información de recetas y aquellos que no la contienen. Una vez así clasificados, se toman solamente los documentos contenedores de recetas y se aplica el segundo módulo. En este, las palabras de cada documento son etiquetadas según su categoría gramatical. Posteriormente, sobre el texto etiquetado se buscan marcadores lingüísticos y se genera un documento de resultado el cual contiene marcados los ingredientes dentro de la receta.

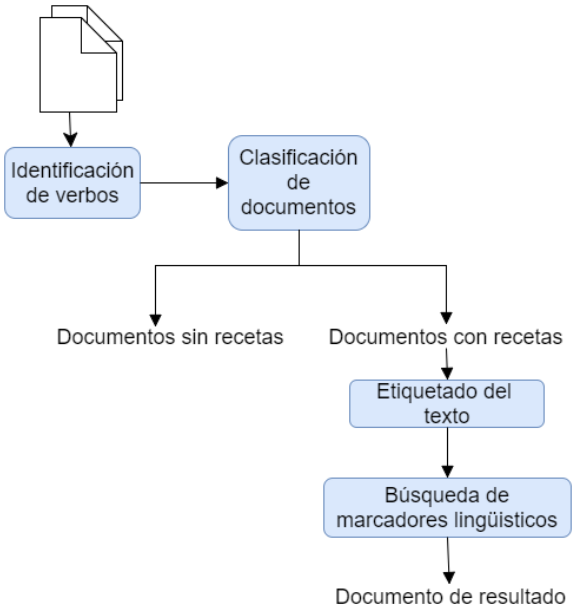


FIGURA 1
Descripción del proceso realizado

Antes de aplicar el procedimiento de dos módulos, se realizó una descarga masiva de

documentos mediante una araña de búsqueda. Esta araña estaba guiada por hipervinculados brindados por los lingüistas del equipo.

En cuanto a las herramientas computacionales utilizadas en el proceso descrito en la figura 1, estas corresponden a agentes automáticos de recolección de información, expresiones regulares y etiquetador de partes del discurso:

- **Agente automático de recolección de información:** Proceso que se ejecuta de forma continua, sigue enlaces y busca adelante por información inferida (cf. Casasola y Gauch 1997).

- **Expresión regular:** Secuencia de caracteres utilizada como patrón para describir, manipular y realizar búsquedas dentro de texto. Es una herramienta sumamente flexible y eficiente para procesamiento de texto (cf. Fitzgerald 2012, Friedl 2006, Habibi 2004).

- **Etiquetador de partes del discurso (POS tagger):** Software para asignar a cada palabra dentro del texto una etiqueta con base en la función que asume en la oración (referida especialmente a la clase léxica o la clase morfológica). Este etiquetado es importante en el área de recuperación de información y procesamiento de lenguaje natural porque encapsula datos propios de la palabra (número, género, tiempo verbal, entre otros), así como de sus palabras vecinas (cf. Hasan, UzZaman y Khan 2007).

2. Procesamiento

2.1. Clasificación de documentos

Este módulo es el encargado de analizar los documentos para clasificarlos en dos categorías: documentos con recetas y documentos sin recetas. La implementación de la lógica del módulo se realizó en dos etapas.

2.1.1. Etapa 1

En esta etapa se utilizaron únicamente documentos de recetas previamente analizados por los lingüistas involucrados en la investigación. Estos documentos se analizaron mediante el uso de un etiquetador de partes del discurso (POS tagger) para identificar de manera automática los verbos presentes en los textos. El resultado del proceso mostraba la lista de los verbos identificados, con su correspondiente forma en infinitivo (lema) y su frecuencia absoluta de aparición en los textos analizados.

A partir de este resultado, se consideró el papel desempeñado por cada verbo en las recetas de cocina. De esta manera, se clasificaron los verbos en dos categorías de significancia (media y alta) basándose en qué tanto es exclusivo cada verbo del dominio de la gastronomía y las recetas de cocina, lo cual es útil para una identificación automática de recetas. Por ejemplo, algunos verbos encontrados en los textos que se estudiaron son de uso común en otros dominios y, por lo tanto, no pueden asociarse de manera única al contexto de cocina. Sin embargo, otros verbos son claramente exclusivos el discurso gastronómico, hecho que permite pensar en la posibilidad de utilizarlos instrumentalmente para la identificación de recetas de cocina dentro de un corpus textual inicialmente indiferenciado.

- **Significancia media**: Verbos comunes en diversos dominios y, por tanto, no exclusivos de los contextos de recetas. Por ejemplo: cocinar, servir, hacer, mezclar.

- **Significancia alta**: Verbos que pueden asociarse comúnmente al contexto de la gastronomía y las recetas de cocina. Por ejemplo: amasar y picar.

2.1.2. Etapa 2

Para la segunda etapa, se tomó como base de la clasificación de documentos el resultado obtenido en la etapa 1 correspondiente a la significancia de los verbos. Además, se procedió a la creación de un agente automático de recolección de información (v. Casasola y Gauch 1997) para la descarga masiva de documentos de internet. Este agente utiliza inicialmente un conjunto de páginas web a las

que se les conoce como *semillas* del agente. Estas semillas son visitadas y cualquier enlace incluido en ellas es agregado a la lista de páginas web por visitar y descargar. Por otro lado, para asegurarse de que la mayoría de los documentos descargados sean gastronómicos, este agente confronta los documentos por descargar contra una lista de páginas web ya validadas. Esta lista de páginas válidas fue previamente brindada por los lingüistas del equipo, quienes las analizaron según criterios de contenido y las valoraron como páginas con gran densidad de recetas de cocina.

Por otro lado, para la clasificación automática de documentos se construyó un programa que recibe como parámetro la ubicación de la carpeta donde se encuentran los documentos descargados por el agente y procede a analizar cada uno de ellos. El primer paso del proceso es el etiquetado del texto de los documentos mediante un POS tagger para identificar los verbos presentes.

Posteriormente, se verifica si los verbos identificados se encuentran clasificados en la lista de verbos de significancia media o alta. Si el documento contiene al menos un verbo de significancia alta o si contiene al menos cuatro verbos de significancia media, el documento es clasificado como contenedor de recetas. Por lo tanto, todo verbo no clasificado previamente tendrá impacto nulo en la clasificación del documento.

Con respecto a la elección de un mínimo de cuatro verbos de la categoría de significancia media, esto se debe a que cada uno de estos verbos por sí solo no permite asegurar que el documento contiene recetas. Sin embargo, la aparición de varios de estos verbos en un mismo documento aumenta la posibilidad de que tal documento efectivamente contenga descripciones culinarias.

Por último, el programa genera un archivo de resultado en el cual se indica la clasificación asignada a cada documento analizado. Este archivo de resultado presenta, además, los verbos detectados en el texto del documento durante el análisis, tal como se ejemplifica en la tabla 1.

TABLA 1

Ejemplos de resultado de clasificación de documentos

RESULTADO
El archivo 0 es receta. Verbos: [cocer] Verbos [comer, cocinar, adornar, servir, probar]
El archivo 102 es NO receta. Verbos: [ ] Verbos: [servir]
El archivo 135 es receta. Verbos: [hornear, batir] Verbos: [ ]
El archivo 179 es receta. Verbos: [ ] Verbos: [moler, majar, arrollar, pelar, enfriar, cocinar, cortar]

Por otro lado, con el fin de evaluar la precisión del programa, se realizó una clasificación manual de los documentos, de modo que se pudieran comparar los resultados automáticos contra los manuales. Según la matriz de confusión resultante (tabla 2), la precisión del sistema resultó ser del 77%.

TABLA 2

Matriz de confusión de la clasificación de documentos: clasificación obtenida contra clasificación real

		Clasificación obtenida	
		Receta	No-Receta
Clasificación real	Receta	301	82
	No-Receta	88	37

2.2. Identificación de ingredientes

Este módulo trabaja únicamente sobre los documentos clasificados en el módulo anterior como documentos con recetas. Antes de iniciar el procesamiento propio de este componente, se aplica sobre el texto un preprocesamiento para normalizarlo y estandarizarlo. La normalización consiste en la eliminación o sustitución de los caracteres especiales, conversión de todo el texto a minúscula, eliminación de espacios múltiples entre palabras, entre otros. Por su parte, la estandarización consiste en la transformación de caracteres especiales de representación de fracciones (por ejemplo, ½, ¼) a su forma normal, es decir, mediante caracteres individuales (por ejemplo, 1/2, 1/4,

respectivamente). Es conveniente señalar que en un inicio este preprocesamiento no se realizaba. Sin embargo, la falta de estandarización en el texto provocaba problemas en la identificación de los marcadores lingüísticos.

Ahora bien, el proceso realizado en este segundo módulo ha tenido dos etapas. La mayor diferencia entre ambas es el paso de usar expresiones regulares basadas en reglas con palabras específicas a expresiones regulares basadas en reglas con categorías gramaticales.

2.2.1. Etapa 1

Los lingüistas del equipo brindaron un corpus de prueba (CP) constituido por texto plano rico en información gastronómica (y, especialmente, denso en recetas). El corpus CP fue extraído de internet por medio de un proceso automatizado y, posteriormente, depurado y prenormalizado de forma masiva para ser utilizado como corpus anónimo. También ofrecieron una lista de formas lingüísticas postuladas como marcadores definicionales de ingredientes –en la figura 2 se pueden ver algunos ejemplos–.

2_barras_de_
2_botellas_de_
2_cabezas_de_
2_[X]_grandes
2_[X]_medianas
2_[X]_pequeñas
2_[X]_picadas
2_[X]_tiernos
2_cucharadas_de_
2_ cucharaditas_de_
2_dientes_de_
2_hojas_de_
2_[X]_batidos
2_[X]_duros

FIGURA 2
Ejemplos de marcadores lingüísticos iniciales

Estos candidatos a marcadores habían sido previamente identificados mediante un proceso manual de análisis de un corpus base (CB) –diferente del corpus de prueba CP ya mencionado–; luego fueron generalizados (o pregeneralizados) utilizando una simbología que pudiera servir de transición hacia la posterior formulación mediante expresiones regulares. La simbología de esa generalización inicial se puede observar en la tabla 3.

TABLA 3
Simbología de los marcadores lingüísticos iniciales

Notación	Significado
[X]	Conjunto obligatorio y variable de 1 o más letras
[[X]]	Conjunto opcional y variable de 1 o más letras
[[elemento]]	Elemento opcional
_	Espacio en blanco

Los marcadores lingüísticos brindados fueron, entonces, transformados por medio de simbología utilizada por las expresiones lingüísticas del lenguaje de programación seleccionado para la automatización. Este paso a expresiones regulares permitió abstraer los patrones iniciales y, por ende, se disminuyó la cantidad de patrones por evaluar. La abstracción se logró en su mayoría al pasar de números específicos (0, 1, 2, 3, 4, 5,... n) a una expresión regular de un conjunto de números, como en:

3_kilos_de
4_kilos_de
} [0-9]+_kilos_de

1/2_taza_de
1/4_taza_de
} [0-9]+/[0-9]+_taza_de

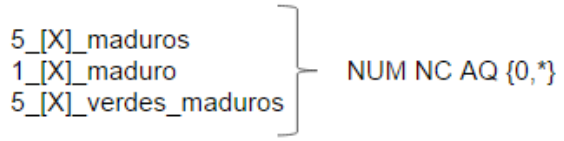
Estas expresiones regulares se buscaron en el texto para identificar los puntos de inserción de ingredientes en las recetas de cocina. Por último, se generaba un documento de resultados en el que se se señalaban los marcadores encontrados y se desplegaba la frecuencia absoluta de aparición de cada uno de ellos. De esta manera se lograron identificar también marcadores que no resultaban útiles para la investigación, debido a su baja o nula frecuencia absoluta de aparición. Además, se

logró identificar los casos de superposición de patrones; una vez identificadas y evaluadas estas superposiciones, se eliminaron las expresiones regulares que producían las redundancias.

A pesar de lograr identificar la mayoría de los ingredientes de cocina por medio de marcadores, también había aquellos que no lograban ser detectados. Una vez analizados los resultados, se observó que muchos de los casos de ingredientes no identificados se debían a falta de coincidencia entre el género o número gramatical de las formas que aparecían en el texto y el género o número de las formas postuladas por los marcadores. Además, no todas las medidas y sus diversas formas de escribirse estaban consideradas en los marcadores. A partir de la revisión de esos resultados, se llegó a la conclusión de que era necesario generalizar de manera un poco diferente los marcadores, de modo que siempre incluyeran al menos todas las posibles inflexiones de género y número.

2.2.2. Etapa 2

Esta segunda versión del módulo se desarrolló con el fin de solucionar el problema de las inflexiones nominales (número) y adjetivales (género y número), además de otras generalizaciones pertinentes. Para esto se empleó un etiquetador de partes del discurso (POS tagger) con un modelo correspondiente al lenguaje español. Este etiquetador permitió pasar de las expresiones regulares de la primera etapa a expresiones regulares basadas en categorías gramaticales.



Para la construcción de estas nuevas expresiones regulares se utilizaron los marcadores brindados inicialmente. Además, las categorías de interés corresponden únicamente a los valores numéricos, sustantivos comunes, adjetivos calificativos, signos de puntuación y preposiciones. Esto permitió disminuir más la cantidad de marcadores por evaluar en el texto, ya que las categorías gramaticales generalizaron valores específicos.

Primeramente, este módulo toma cada uno de los documentos y etiqueta su texto según las partes del discurso. En el siguiente paso, se analiza el texto etiquetado para identificar la presencia de los marcadores definicionales de interés. Por último, el proceso genera un documento de resultados por cada documento analizado. En este documento de resultados se presentan señalados los marcadores definicionales e ingredientes encontrados en el documento.

En cuanto a los resultados, estos se evaluaron cuantitativamente según la cantidad de ingredientes identificados correctamente con respecto al total de los ingredientes en los documentos analizados. Este módulo identificó correctamente y en forma automática el 53% de los ingredientes.

3. Conclusiones

El uso de herramientas de computación para el procesamiento automático de texto demostró ser de utilidad para la recolección, clasificación automática e identificación de ingredientes de recetas de gastronomía con base en marcadores lingüísticos. En relación con el POS tagger, este permitió identificar los verbos dentro del texto para hacer una discriminación automática de documentos, así como considerar las inflexiones nominales y adjetivales en los marcadores lingüísticos. Además, el uso de esta herramienta permitió disminuir la cantidad de tiempo en el desarrollo del proceso, ya que no se requirió realizar manualmente la clasificación en categorías gramaticales de cada una de las palabras en los documentos.

Asimismo, la combinación entre expresiones regulares y categorías gramaticales permitió la generalización y expansión de los marcadores que los lingüistas del equipo habían brindado pregeneralizados.

Trabajo futuro. La investigación presentada en este artículo proyecta extenderse y optimizarse para la obtención de mejores resultados. El plan es dividir las tareas computacionales en dos temas a ser desarrollados como trabajos finales de

investigación aplicada (TFIA) en la Maestría en Computación. El primero se refiere a la clasificación automática de documentos utilizando aprendizaje de máquina y marcadores lingüísticos. Este trabajo se enfocará en la clasificación de textos para diferenciar entre archivos que tienen información gastronómica (específicamente, recetas) y archivos que no la tienen; esto requerirá trabajar en conjunto con los expertos en el área de lingüística para brindar pesos a los verbos utilizados para la clasificación.

Por otro lado, el segundo tema corresponde al análisis automático de textos de recetas de cocina para la identificación de procesos paralelos y secuenciales. Este trabajo podrá utilizar como base el proceso realizado para la identificación de los ingredientes en las recetas de cocina por medio de patrones definicionales, y requerirá además contar con un listado de marcadores lingüísticos definitorios asociados a los diversos pasos/tareas/etapas de los procedimientos culinarios.

Notas

1. Precisamente por la búsqueda de modelos metodológicos eficaces con vista en los objetivos del análisis de CD, parte importante de los estudios enmarcados en esta línea de investigación es la propuesta y continuo afinamiento de los procedimientos aplicados a las diversas etapas.

Referencias

Alarcón, Rodrigo. (2003). Análisis lingüístico de contextos definitorios en textos de especialidad. Tesis de licenciatura: Universidad Nacional Autónoma de México.

Alcina, Amparo y Esperanza Valero. (2008): "Análisis de las definiciones del diccionario cerámico científico-práctico. Sugerencias para la elaboración de patrones de definición". En: *Debate Terminológico*, 4. http://seer.ufrgs.br/index.php/riterm/article/download/23841/13830. Consulta: 20/02/2017.

Casasola, Édgar y Susan Gauch. (1997). Intelligent Information Agents for the World Wide Web. [Information and Telecommunication Technology Center, Technical report ITTC-FY97-111100-1]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.3628&rep=rep1&type=pdf. Consulta: 22-02-2017.

Elleithy, Khaled (ed.). (2007). *Advances and Innovations in Systems, Computing Sciences and Software Engineering*. Dordrecht, Holanda: Springer.

Fitzgerald, Michael. (2012). *Introducing Regular Expressions*. California: O'Reilly Media Inc.

Friedl, Jeffrey E. F. (2006). *Mastering Regular Expressions* (3a. ed.). California: O'Reilly Media Inc.

Habibi, Mehran. (2004). *Java Regular Expressions: Taming the java.util.regex Engine*. Nueva York: Apress Media, LLC.

Hasan, Fahim Muhammad, Naushad UzZaman y Mumit Khan. (2007). "Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla". En: Elleithy (ed.): 121-126.

Sierra, Gerardo. (2009). "Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos". En: *linguaMATICA*, 2: 13-37.

Sierra, Gerardo y Rodrigo Alarcón. (2002). "Identification of recurrent patterns to extract definitory contexts". En: *Lecture Notes in Computer Science*, 2276: 436-438.

Sierra, Gerardo, Rodrigo Alarcón y César Aguilar. (2006). "Extracción automática

de contextos definitorios en textos especializados". En: *Revista de Procesamiento de Lenguaje Natural*, 37: 351-352.

Sierra, Gerardo, Mara Pozzi y Juan Manuel Torres (eds.). (2009). Proceedings. (1st) International Workshop on Definition Extraction, 18 de setiembre de 2009, Borovets, Bulgaria. https://aclweb.org/anthology/W/W09/W09-4400.pdf. Consulta: 20/02/2017.

Soler, Victoria. (2005). Patrones lingüísticos para la búsqueda de información conceptual en el corpus textual especializado de la cerámica TXTCera. http://repositori.uji.es/xmlui/bitstream/handle/10234/79115/forum_2004_50.pdf?sequence=1. Consulta: 20/02/2017.

Valero, Esperanza. (2009). Los marcadores lingüísticos en las definiciones del grupo conceptual 'procesos de fabricación cerámica'. http://repositori.uji.es/xmlui/bitstream/handle/10234/78051/forum_2008_22.pdf?sequence=1. Consulta: 20/02/2017.

Valero, Esperanza y Amparo Alcina. (2009). "Linguistic realization of conceptual features in terminographic dictionary definitions". En: Sierra, Pozzi y Torres (eds.): 54–60.