

# PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

## FACULTAD DE CIENCIAS E INGENIERÍA



PONTIFICIA  
**UNIVERSIDAD**  
**CATÓLICA**  
DEL PERÚ

### IMPLEMENTACION DE UN SOFTWARE DE APOYO A LA ESCRITURA DE RESÚMENES DE TEXTOS CIENTÍFICOS EN ESPAÑOL

Tesis para optar el Título de **Ingeniero Informático**, que presenta el bachiller:

**Irvin Rosendo Vargas Campos**

**ASESOR: Ing. Fernando Alva Manchego**

Lima, diciembre de 2013

## RESUMEN

Desde hace tiempo se viene comentando que los estudiantes universitarios presentan serios problemas de expresión escrita. En diversas fuentes de información, tales como artículos de investigación científica, tesis, u otros medios académicos y profesionales, se puede apreciar diversos errores de redacción. Ésta es una situación que se considera inadmisibles en personas con un alto nivel de instrucción formal, especialmente porque todas ellas ya han pasado alrededor de once años de escolarización en la que aprobaron diversas materias relativas a la enseñanza de su lengua materna.

Como medida para solucionar este problema, se busca promover la enseñanza de la organización de las ideas. Existen varias técnicas que ayudan a organizar las ideas y preparar la información antes de la redacción del ensayo, monografía o artículo científico. Una de las técnicas más básicas es la redacción del resumen.

Se sabe que la redacción del resumen de los textos científicos es una técnica básica y fundamental para la organización de ideas y preparación de información para redactar correctamente textos científicos más complejos. Por tal motivo, el presente proyecto de fin de carrera presenta la implementación de un software de apoyo a la escritura de resúmenes de textos científicos en español, el cual ayudará al escritor a redactar resúmenes de sus textos científicos con una estructura adecuada.

Para poder llevarlo a cabo, primero se formó un corpus de 44 resúmenes de textos científicos en español, que sirven para el entrenamiento y prueba del modelo clasificador AZEsp. Para formar el corpus, se tuvo como estructura óptima de los textos la presencia de 6 categorías: Contexto, Brecha, Propósito, Metodología, Resultado y Conclusión.

Luego, se procedió a determinar un conjunto de 7 características (atributos), las cuales serían utilizadas para identificar cada una de las categorías. Posteriormente, se implementaron una serie de algoritmos para la extracción de los valores de dichos atributos de cada oración de los resúmenes de textos científicos para que sean utilizadas por el modelo. Una vez obtenidos dichos valores, éstos fueron utilizados para la implementación del modelo clasificador AZEsp y evaluación de su desempeño utilizando métricas tales como Precision, Recall y F-Measure.

Finalmente, se implementó el ambiente de ayuda SciEsp, el cual utiliza el modelo clasificador AZEsp para clasificar automáticamente las oraciones de los resúmenes de textos científicos en español ingresados por el usuario, siguiendo una estructura predefinida.

Se hizo una serie de experimentos para evaluar el desempeño del modelo clasificador AZEsp. Se obtuvo diferentes resultados; sin embargo, el más resaltante fue que el modelo logró un desempeño de 65.4%. Esto demuestra que la herramienta informática propuesta (SciEsp) está apta para su utilización. En conclusión, los estudiantes universitarios podrán emplear esta herramienta para la redacción de sus resúmenes; ellos podrán identificar sus errores y deficiencias en la redacción, y serán capaces de mejorar de forma autodidacta.

FACULTAD DE  
CIENCIAS E  
INGENIERÍA  
ESPECIALIDAD DE  
INGENIERÍA INFORMÁTICA



PONTIFICIA  
UNIVERSIDAD  
CATÓLICA  
DEL PERÚ

**TEMA DE TESIS PARA OPTAR EL TÍTULO DE INGENIERO INFORMÁTICO**

**TÍTULO:** IMPLEMENTACION DE UN SOFTWARE DE APOYO A LA ESCRITURA DE RESÚMENES DE TEXTOS CIENTÍFICOS EN ESPAÑOL.

**ÁREA:** CIENCIAS DE LA COMPUTACIÓN # 514

**PROPONENTE:** Alva Manchego, Fernando

**ASESOR:** Alva Manchego, Fernando

**ALUMNO:** Vargas Campos, Irvin Rosendo

**CÓDIGO:** 20080249

**TEMA N°:** \_\_\_\_\_

**FECHA:** San Miguel, 16 de noviembre de 2013




**DESCRIPCIÓN**

Desde hace tiempo se viene comentando que los estudiantes universitarios presentan serios problemas de expresión escrita. En diversas fuentes de información, tales como artículos de investigación científica, tesis, u otros medios académicos y profesionales, se puede apreciar diversos errores de redacción. Ésta es una situación que se considera inadmisibles en personas con un alto nivel de instrucción formal, especialmente porque todas ellas ya han pasado alrededor de once años de escolarización en la que aprobaron diversas materias relativas a la enseñanza de su lengua materna.

Como medida para solucionar este problema, se busca promover la enseñanza de la organización de las ideas. Existen varias técnicas que ayudan a organizar las ideas y preparar la información antes de la redacción del ensayo, monografía o artículo científico. Una de las técnicas más básicas es la redacción del resumen.

Se sabe que la redacción del resumen de los textos científicos es una técnica básica y fundamental para la organización de ideas y preparación de información para redactar correctamente textos científicos más complejos. Por tal motivo, el presente proyecto de fin de carrera presenta el desarrollo de una herramienta informática de apoyo a la escritura de resúmenes de textos científicos en español, el cual ayudará al escritor a redactar resúmenes de sus textos científicos con una estructura adecuada.

**OBJETIVO GENERAL**

 Desarrollar una herramienta informática de apoyo a la escritura de resúmenes de textos científicos en español que utilice modelos de aprendizaje de máquina supervisado.

Av. Universitaria 1801  
San Miguel, Lima - Perú

Apartado Postal 1761  
Lima 100 - Perú

Teléfono:  
(511) 626 2000 Anexo 4801

FACULTAD DE  
**CIENCIAS E  
INGENIERÍA**  
ESPECIALIDAD DE  
INGENIERÍA INFORMÁTICA



PONTIFICIA  
**UNIVERSIDAD  
CATÓLICA**  
DEL PERÚ

### OBJETIVOS ESPECÍFICOS

1. Formar un corpus de resúmenes de textos científicos en español que sirvan para el entrenamiento y prueba de los modelos de aprendizaje.
2. Determinar las características (atributos) que serán extraídas de cada oración de los resúmenes de textos científicos para que sean utilizadas por los modelos de aprendizaje.
3. Implementar un modelo clasificador y evaluar su desempeño utilizando métricas tales como Precision, Recall y F-Measure.
4. Implementar una aplicación que clasifique automáticamente las oraciones de los resúmenes de textos científicos en español ingresados por el usuario, siguiendo una estructura pre-definida.

### ALCANCE

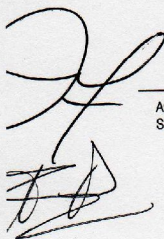
El proyecto de fin de carrera se relaciona con la capacidad de redacción del sector universitario, específicamente la redacción de resúmenes de textos científicos. Se ha elegido este tema en especial debido a la gran importancia que tiene saber redactar en el campo social, académico y laboral en la actualidad.

Se trata de un proyecto de Procesamiento de Lenguaje Natural, rama de las Ciencias de la Computación. El proyecto se basa en el desarrollo de una herramienta informática que ayude en la redacción de resúmenes de textos científicos en español. Se optó por enfocar el proyecto en los textos en español, no sólo por el dominio que se tiene sobre este idioma, sino también debido a que no existen muchas herramientas de ayuda en la redacción de estos textos.

En cuanto al aprendizaje de la herramienta, se tendrá un corpus de estudio que estará compuesto por una gama, no mayor a 50, de resúmenes de textos científicos en español. Cabe resaltar que éstos serán previamente revisados con el fin de garantizar su correcta redacción, y por ende un mejor funcionamiento de la herramienta en cuestión.

La herramienta será capaz de identificar los diversos componentes que conforman la estructura de los resúmenes de textos científicos ingresados por el usuario.

Con el desarrollo de esta herramienta, se busca que los universitarios puedan ver sus deficiencias en el tema y así mejorar su redacción de resúmenes de textos científicos de forma autodidacta y eficiente. Por otro lado, se desea, si es posible, poder llegar a ser un punto de partida para otros proyectos del mismo rubro con un alcance más profundo.



Av. Universitaria 1801  
San Miguel, Lima – Perú

Apartado Postal 1761  
Lima 100 – Perú

Teléfono:  
(511) 626 2000 Anexo 4801



## Dedicatoria

Dedico la presente tesis, en primer lugar, a Dios por mostrarme, día a día, que con humildad, paciencia y sabiduría todo es posible.

A mis padres y hermano por su apoyo y comprensión incondicional a lo largo de toda mi vida universitaria.

Y a todas aquellas personas muy allegadas que siempre tuvieron una palabra de aliento en los momentos difíciles y que han sido muy importantes en mi vida.

## Agradecimientos

Agradezco a Dios, mi familia y todas las personas por la paciencia y el apoyo incondicional para la realización con éxito de mi proyecto de fin de carrera.

Y un agradecimiento muy especial al Ing. Fernando Alva Manchego por su paciencia, apoyo y asesoría constante durante el desarrollo del presente proyecto.

## Tabla de contenido

RESUMEN	2
ÍNDICE DE FIGURAS	9
ÍNDICE DE TABLAS	10
<b>CAPÍTULO 1</b>	<b>11</b>
1 INTRODUCCIÓN	11
2 PROBLEMÁTICA	11
3 MARCO TEÓRICO	13
3.1 CONCEPTOS RELACIONADOS AL PROBLEMA	13
3.2 CONCEPTOS RELACIONADOS A LA PROPUESTA DE SOLUCIÓN	16
4 ESTADO DEL ARTE	23
4.1 FORMAS EXACTAS DE RESOLVER EL PROBLEMA	23
4.2 PRODUCTOS COMERCIALES PARA RESOLVER EL PROBLEMA	24
4.3 PRODUCTOS NO COMERCIALES (DE INVESTIGACIÓN) PARA RESOLVER EL PROBLEMA	25
4.4 CONCLUSIONES SOBRE EL ESTADO DEL ARTE	31
<b>CAPÍTULO 2</b>	<b>33</b>
1 OBJETIVO GENERAL	33
2 OBJETIVOS ESPECÍFICOS	33
3 RESULTADOS ESPERADOS	33
4 HERRAMIENTAS, MÉTODOS Y PROCEDIMIENTOS	33
4.1 MAPEO	34
4.2 HERRAMIENTAS, MÉTODOS Y PROCEDIMIENTOS	34
4.3 METODOLOGÍAS	37
5 ALCANCE	38
5.1 LIMITACIONES	38
5.2 RIESGOS	38
6 JUSTIFICACIÓN Y VIABILIDAD	39
6.1 JUSTIFICATIVA DEL PROYECTO DE TESIS	39
6.2 ANÁLISIS DE VIABILIDAD DEL PROYECTO DE TESIS	39
7 PLAN DE ACTIVIDADES	40
<b>CAPÍTULO 3: CORPUS DE RESÚMENES</b>	<b>42</b>
1 BREVE DESCRIPCIÓN	42
2 CORPUS VS ÁREAS DE INFORMÁTICA	42
3 NÚMERO DE COMPONENTES POR RESUMEN	43
4 FRECUENCIA DE LOS COMPONENTES EN EL CORPUS	43
<b>CAPÍTULO 4: CONJUNTO DE ATRIBUTOS QUE SE EMPLEARÁN PARA LA IDENTIFICACIÓN DE LAS CATEGORÍAS</b>	<b>45</b>
1 BREVE DESCRIPCIÓN	45
2 VISIÓN GENERAL DE LOS ATRIBUTOS	45
3 DESCRIPCIÓN DETALLADA DE LOS ATRIBUTOS	46
3.1 TAMAÑO	46
3.2 LOCALIZACIÓN	46
3.3 EXPRESIÓN	46
3.4 TIEMPO, VOZ Y MODAL	47
3.5 HISTÓRICO	48
<b>CAPÍTULO 5: EXTRACCIÓN DE LOS VALORES DE LOS ATRIBUTOS</b>	<b>49</b>
1 BREVE DESCRIPCIÓN	49
2 PROCESO DE DETERMINACIÓN DE ATRIBUTOS	49
2.1 TOKENIZACIÓN Y DELIMITACIÓN DE ORACIONES	49
2.2 IDENTIFICACIÓN DE EXPRESIONES	50

2.3	POS-TAGGING	51
2.4	PROCESAMIENTO SINTÁCTICO	51
CAPÍTULO 6: EVALUACIÓN DEL MODELO CLASIFICADOR		54
1	BREVE DESCRIPCIÓN DEL CLASIFICADOR ESTADÍSTICO	54
2	EVALUACIÓN DEL CLASIFICADOR	54
3	CONSIDERACIONES FINALES	57
CAPÍTULO 7: CONCLUSIONES Y TRABAJOS FUTUROS		58
REFERENCIAS BIBLIOGRÁFICAS		60





## Índice de figuras

FIGURA 1: MARCO DE CLASIFICACIÓN SUPERVISADA. BASADO EN UN GRÁFICO DE NLTK, S/F. ....	20
FIGURA 2: ORGANIZACIÓN DEL CORPUS DE DATOS PARA EL ENTRENAMIENTO DE CLASIFICADORES SUPERVISADOS. FUENTE: NLTK, S/F. ....	21
FIGURA 3: SAMPLE FEEDBACK SCREEN OF IEA. FUENTE: RUDNER ET AL., 2000. ....	26
FIGURA 4: SCORING AND RECALCULATION WINDOW IN BETSY. FUENTE: RUDNER ET AL., 2000. ....	27
FIGURA 5: EVALUACIÓN DE RESÚMENES EN PORTUGUÉS. FUENTE: SciPo, 2000. ....	28
FIGURA 6: EVALUACIÓN DE INTRODUCCIONES EN PORTUGUÉS. FUENTE: SciPo, 2000. ....	29
FIGURA 7: EVALUACIÓN DE RESÚMENES EN INGLÉS. FUENTE: SciPo-FARMACIA, 2000. ....	30
FIGURA 8: EL MODO DE CLASIFICACIÓN. FUENTE: WEKA 3, S/F. ....	35
FIGURA 9: DISTRIBUCIÓN DE LAS CATEGORÍAS EN EL CORPUS. ....	44
FIGURA 10: ETAPAS DEL PROCESO DE DETERMINACIÓN DE ATRIBUTOS. ....	49
FIGURA 11: EJEMPLO DE VERBOS SIMPLES Y COMPUESTOS. FUENTE: WIKIPEDIA, 2013. ....	53
FIGURA 12: RESULTADOS DEL EXPERIMENTO 1, OBTENIDOS POR WEKA. ....	55
FIGURA 13: RESULTADOS DEL EXPERIMENTO 2, OBTENIDOS POR WEKA. ....	56
FIGURA 14: RESULTADOS DEL EXPERIMENTO 3, OBTENIDOS POR WEKA. ....	57



## Índice de tablas

TABLA 1: DOCENTES DE PRIMER Y SEGUNDO GRADO DE PRIMARIA QUE PARTICIPARON EN EL PROGRAMA DE CAPACITACIÓN BÁSICA Y ESPECIALIZADA. ....	12
TABLA 2: TIPOS DE REDACCIÓN. FUENTE: MOLESTINA ET AL., 1988.....	15
TABLA 3: TIPOS DE TEXTOS CIENTÍFICOS. FUENTE: MOLESTINA ET AL., 1988; UNESCO, 1951. ....	16
TABLA 4: NIVELES DE CONOCIMIENTO EN EL PLN. ....	17
TABLA 5: VERDADEROS Y FALSOS POSITIVOS Y NEGATIVOS. ....	22
TABLA 6: ESTRUCTURA IDEAL DE UN RESUMEN DE TEXTO CIENTÍFICO. FUENTE: SCIPo, 2000. ....	24
TABLA 7: CARACTERÍSTICAS DE LOS SOFTWARE DEDICADOS A LA EVALUACIÓN DE LA ESCRITURA. ....	31
TABLA 8: MAPEO DE RESULTADOS ESPERADOS Y HERRAMIENTAS, MÉTODOS Y PROCEDIMIENTOS A USARSE. ....	34
TABLA 9: SERVICIOS DE ANÁLISIS LINGÜÍSTICOS DISPONIBLES PARA CADA LENGUA. FUENTE: PADRÓ & STANILOVSKY, 2012. ....	36
TABLA 10: RIESGOS DEL PROYECTO DE FIN DE CARRERA.....	39
TABLA 11: PLAN DE ACTIVIDADES QUE SE REALIZARÁN A LO LARGO DEL PROYECTO DE FIN DE CARRERA. ....	41
TABLA 12: DISTRIBUCIÓN DEL CORPUS A TRAVÉS DE LAS ÁREAS DE INFORMÁTICA. ....	42
TABLA 13: NÚMERO DE COMPONENTES POR RESUMEN.....	43
TABLA 14: ESQUEMA DE CLASIFICACIÓN. ....	45
TABLA 15: RESUMEN DEL CONJUNTO DE ATRIBUTOS. ....	46
TABLA 16: EJEMPLOS DE EXPRESIONES ESTÁNDARES.....	47
TABLA 17: EJEMPLOS DE EXPRESIONES ESTÁNDARES ENCONTRADAS EN EL CORPUS DE RESÚMENES CIENTÍFICOS. ....	50
TABLA 18: CATEGORÍAS DE LA ETIQUETACIÓN DE LOS VERBOS. FUENTE: FREELING, S/F. ....	52
TABLA 19: EJEMPLO DE ETIQUETACIÓN DE LOS VERBOS. FUENTE: FREELING, S/F.....	52



## CAPÍTULO 1

### 1 Introducción

El presente proyecto de fin de carrera tiene como objetivo desarrollar una herramienta informática de apoyo a la escritura de resúmenes de textos científicos en español. Esta herramienta beneficiará a los universitarios e incluso estudiantes de nivel escolar u otro que estén interesados en el tema de la redacción de resúmenes de textos científicos. Gracias a esta herramienta, ellos podrán identificar sus errores y deficiencias en la redacción, y serán capaces de mejorarla de forma autodidacta.

En este capítulo, se planteará un análisis a la problemática existente relacionada a la redacción en el Perú, y se describirá brevemente la propuesta de solución a este problema. Además, se tendrá una sección dedicada al marco teórico, donde se mostrarán los principales conceptos relacionados al problema, así como los relacionados a la solución. Finalmente, en la sección del estado del arte, se presentarán una serie de herramientas utilizadas hasta el momento para resolver este problema, ya sea total o parcialmente.

### 2 Problemática

Desde hace tiempo se viene comentando que los estudiantes universitarios presentan serios problemas de expresión escrita. En diversas fuentes de información, tales como artículos de investigación científica, tesis, u otros medios académicos y profesionales, se puede apreciar diversos errores de redacción. Ésta es una situación que se considera inadmisibles en personas con un alto nivel de instrucción formal, especialmente porque todas ellas ya han pasado alrededor de once años de escolarización en la que aprobaron diversas materias relativas a la enseñanza de su lengua materna (SANCHEZ, 2005).

En el Perú, la mala redacción de textos es una realidad que no pasa desapercibida. Los universitarios presentan este problema debido a su inadecuada formación en la escuela primaria y secundaria en el ámbito de la redacción. Este hecho promueve que los alumnos tengan dificultades para emplear la redacción como forma personal de procesar información y como una herramienta para interactuar con su entorno (RAMOS, 2011).

Un factor importante que promueve la mala redacción es la falta de interés de los profesores en lo que respecta a la capacitación (básica o especializada) en la enseñanza de los niños. El Ministerio de Educación (MINEDU) y el Instituto Nacional de Estadística e Informática (INEI) muestran las siguientes estadísticas sobre la capacitación de los profesores de primer y segundo grado de primaria.

Tipo de Capacitación	Indicador (%)	Denominador del indicador
Básica	32.1	Número total de docentes de primer y segundo grado de Educación Primaria
Especializada	12.4	Número total de docentes de primer y segundo grado de Educación Primaria que han participado en la capacitación básica

**Tabla 1: Docentes de primer y segundo grado de primaria que participaron en el programa de capacitación básica y especializada.**

Como se puede ver en las estadísticas mostradas en la Tabla 1, los profesores no se encuentran adecuadamente capacitados debido a su inasistencia a los programas brindados. Esto afecta directamente a la enseñanza de los alumnos en diferentes áreas porque “todo lo que el alumno puede ‘aprender’ es mérito o demérito del profesor y de la escuela” (RAMOS, 2011). Una de las disciplinas afectadas es la redacción de textos ya que “el maestro debe ser quien combate, dentro de las escuelas, todos los vicios del lenguaje o barbarismos” (HIRSH & LIMO, 2006).

Como se pudo apreciar anteriormente, la mala redacción empieza a formarse desde la escuela primaria, hasta incluso antes. Entonces, ¿qué pasa con los alumnos universitarios, quienes ya pasaron por la etapa escolar? Si ellos recibieron una mala enseñanza en lo que concierne a la escritura de textos, no tienen la oportunidad de retroceder el tiempo para aprender bien lo que no pudieron. Como medida para solucionar esto, algunas universidades como la PUCP, UPC, entre otras, han incorporado cursos de redacción o argumentación en sus primeros ciclos. Por ejemplo, según lo revisado en la página web de Estudios Generales de Ciencias de la PUCP, en el plan de estudios se cuenta con los cursos de “Introducción a la Comunicación Oral y Escrita”, “Redacción y Comunicación”, “Lengua y Composición”, etc. Se puede decir que al confirmar que los alumnos no pueden formalizar sus pensamientos en un texto, las universidades se han visto obligadas a establecer asignaturas escolares en su currículo, que ocupan espacios en los que deberían ofrecerse otros cursos de la propia carrera (RAMOS, 2011).

Como se puede ver, los estudiantes universitarios deben afrontar diversos problemas en cuanto al aprendizaje de la redacción en general. Con la solución dada anteriormente, se puede decir que el estudiante podrá comprender definiciones generales y conocer buenas prácticas en la redacción, lo cual implica un gran avance; sin embargo, el mayor desafío se encuentra más adelante, éste radica en el aprendizaje de la redacción de textos científicos.

Como bien se sabe, en la etapa de formación académica, se les ha asignado a múltiples estudiantes universitarios escribir una diversidad de textos, en los cuales representaban sus ideas, conocimientos y opiniones. Pocos de estos escritos les exigieron ir más allá de su caudal de información y experiencia. Sin embargo, a medida que avanzan académicamente, ya se comienzan a encontrar frente a la necesidad de ampliar su conocimiento, visión u horizonte; sobre todo cuando se detecta la posibilidad de explorar una idea, de probar algún criterio, de resolver cierto problema, de poseer más información o cuando requieren elaborar argumentos sólidos que vayan en su ayuda (HUAMÁN, s/f). Cuando la dificultad se vuelve insuperable, se ven en la necesidad de investigar, de utilizar nuevos materiales y de ir más allá de sus iniciales recursos personales. Los resultados de dicho proceso de investigación se presentan en una tesis, una monografía, un informe, o mejor dicho un “texto científico”.

Los textos científicos tienen un estilo propio y una estructura definida. Por ejemplo, según SLAFER (2009), el artículo científico tiene como estructura: título, autoría y afiliación, resumen y palabras clave, introducción, material y métodos, resultados, discusión y conclusiones, reconocimientos, y referencias bibliográficas. El desconocimiento o el poco entendimiento, por parte de los estudiantes universitarios, de dichas estructuras generan problemas en la redacción de este tipo de texto. Como resultado del análisis anterior, se puede concluir que el aprendizaje de la redacción de textos científicos es un desafío especial, pues el interesado enfrentará diversas dificultades teóricas si desea dominar este género de escritura.

Haciendo un análisis general, se concluye que a pesar de que existan cursos en la universidad que te enseñen cómo redactar, estos cursos no son suficientes para poder aprender a redactar textos científicos. Para lograrlo, es necesario conocer la estructura del texto, como se había mencionado anteriormente, y esto implica saber organizar las ideas para poder construir dicha estructura.

La organización de las ideas es una operación más crítica que el acopio de las mismas; ésta requiere la utilización de mecanismos asociativos capaces de captar similitudes, construir razonamientos (distinguiendo entre premisas y conclusiones) y desarrollar tesis coherentes (HUAMÁN, s/f). Existen varias técnicas que ayudan a organizar las ideas y preparar la información antes de la redacción del ensayo, monografía o artículo científico. Una de las técnicas más básicas es la redacción del resumen, ya que al escribirlo solo se utilizan los bloques de información más importantes, no solo se trata de seleccionar hechos o datos, sino de hacer un juicio crítico o una valoración y ello supone la comprensión previa del material (HUAMÁN, s/f).

Teniendo la redacción del resumen de los textos científicos como técnica básica y fundamental para la organización de ideas y preparación de información para redactar correctamente textos científicos más complejos, se ha pensado en implementar un software que ayude al escritor a redactar resúmenes de sus textos científicos con una estructura adecuada. La herramienta evaluará el texto hecho por el escritor e identificará los diversos componentes que conforman la estructura de dicho texto. La estructura óptima está conformada por tres componentes principales (propósito, metodología y resultado), y tres componentes opcionales (contexto, brecha y conclusión) siguiendo las recomendaciones de la Dra. Valeria Feltrim (FELTRIM et al., 2003). Los componentes identificados podrán ser visualizados por el escritor, de tal forma que pueda ver sus deficiencias en la escritura y mejorar su redacción de forma autodidacta.

### **3 Marco teórico**

En esta sección se definirán conceptos que permitirán tener un mejor entendimiento del problema presentado anteriormente. Además, se definirán conceptos necesarios para comprender el producto final, así como el área de informática a la que pertenece.

#### **3.1 Conceptos relacionados al problema**

Para comprender mejor el problema descrito es necesario tener en cuenta los siguientes conceptos:

➤ **Redacción de textos científicos**

Para un mejor entendimiento del tema es necesario conocer la definición de redactar y texto científico.

- ✓ **Redactar:** “Poner por escrito algo sucedido, acordado o pensado con anterioridad” (RAE, 2012).
- ✓ **Texto científico:** “Es aquel cuyo contexto contiene de forma confiable todo el proceso que se requiere en una investigación científica. Su objetivo es comunicar el conocimiento y corresponde a la función referencial de la lengua” (COMPARÁN et al., 2007). La función referencial de la lengua se produce cuando se desea transmitir un mensaje, cuya objetividad y claridad cumplan con el objetivo de informar (COMPARÁN et al., 2007).

Una vez entendido estos conceptos, se procederá a establecer el lugar que le corresponde a la redacción científica dentro del marco general de producción literaria (Ver Tabla 2).

Tipos de redacción	Características
<p>PROSA EMOTIVA, de propaganda</p>	<ol style="list-style-type: none"> <li>1. Contiene poca información.</li> <li>2. Llega a los sentimientos: deseo de exclusividad, amor al lujo, etc.</li> <li>3. Usa palabras emotivas: opulento, aristocrático, belleza, distinción, etc.</li> <li>4. Exagera la verdad.</li> <li>5. Está motivada por un deseo de ganancias.</li> <li>6. No es sistemática: no hay sucesión lógica de ideas.</li> <li>7. Parece que no es sincera.</li> <li>8. Usa recursos tipográficos para dar énfasis: mayúsculas, cursiva, oraciones fragmentarias, párrafos cortos.</li> </ol>
<p>PROSA PERSUASIVA, de propaganda</p>	<ol style="list-style-type: none"> <li>1. Presenta algo de información.</li> <li>2. Hace juicios sin ninguna base.</li> <li>3. Es básicamente persuasiva.</li> <li>4. Trata de influir en la actitud del lector.</li> <li>5. Evita la exageración y la insinceridad.</li> <li>6. Presenta una secuencia lógica de ideas.</li> <li>7. Usa palabras moderadamente emotivas: mejoramiento, mejor servicio, grandes cualidades, entusiasmo, etc.</li> </ol>
<p>DESCRIPCIÓN: imaginativa, subjetiva</p>	<ol style="list-style-type: none"> <li>1. Parte informativa, parte imaginativa y subjetiva.</li> <li>2. Subjetiva en el uso de: yo sentí, me convenció, etc.</li> <li>3. Parece sincera y verdadera.</li> <li>4. Describe principalmente el ánimo del escritor.</li> <li>5. Incluye impresiones específicas de los sentidos: la libélula, el sonido de las alas, los escombros en el bote, etc.</li> <li>6. Usa lenguaje figurativo: los dos años como un espejismo, las olas dando sopapos a la quijada del bote, etc.</li> <li>7. Usa un estilo natural familiar, vocabulario simple.</li> </ol>
<p>CRÍTICA: juicio sin apoyo</p>	<ol style="list-style-type: none"> <li>1. No presenta información específica.</li> <li>2. Está hecha de generalizaciones críticas sin evidencia que los apoye.</li> <li>3. Parece sin prejuicios; incluye tanto juicios favorables como desfavorables.</li> <li>4. Es principalmente seria en tono y lenguaje.</li> <li>5. Incluye afirmaciones subjetivas personales.</li> <li>6. Usa términos críticos levemente técnicos: jerga, barroco.</li> <li>7. Está dirigida al lector con conocimientos científicos básicos.</li> </ol>

<p>PROSA CIENTÍFICA: técnica, general</p>	<ol style="list-style-type: none"> <li>1. Es totalmente informativa.</li> <li>2. Usa términos técnicos sin definirlos.</li> <li>3. Es desinteresada y sincera.</li> <li>4. No incluye juicios, pero hace generalizaciones.</li> <li>5. Es principalmente concreta.</li> <li>6. Es seria en tono y orden.</li> <li>7. No tiene atracción emotiva.</li> <li>8. Está dirigida al lector con conocimientos técnicos.</li> </ol>
<p>PROSA CIENTÍFICA: abstracta, seria</p>	<ol style="list-style-type: none"> <li>1. Es abstracta y general.</li> <li>2. Debe ser informativa.</li> <li>3. No es técnica.</li> <li>4. Es desinteresada y sincera.</li> <li>5. Incluye algunas opiniones bien informadas, sin apoyo.</li> <li>6. Es de tono y lenguaje serio.</li> <li>7. No tiene atracción emotiva.</li> <li>8. Su contenido y vocabulario es popular.</li> </ol>
<p>ESCRITOS CIENTÍFICOS: específicos, históricos</p>	<ol style="list-style-type: none"> <li>1. Es totalmente informativa.</li> <li>2. Basada en fuentes históricas.</li> <li>3. No tiene atracción emotiva.</li> <li>4. Es desinteresada y sincera.</li> <li>5. No incluye juicios sobre el valor.</li> <li>6. Es concreta y específica.</li> <li>7. Es semitécnica.</li> <li>8. Es de lenguaje y orden serios.</li> </ol>

**Tabla 2: Tipos de redacción. Fuente: MOLESTINA et al., 1988.**

De estos ejemplos podemos resumir las siguientes características que tipifican a la literatura científica:

1. Presenta hechos.
2. Es exacta y verdadera.
3. Es desinteresada.
4. Es sistemática.
5. No es emotiva.
6. Excluye opiniones no fundadas.
7. Es sincera.
8. No es argumentativa (deja que los hechos hablen por sí solos).
9. No es directamente persuasiva.
10. No exagera.

Otro punto importante que no se debe pasar por alto es tener claro cuáles son los diferentes tipos de escritos científicos. Éstos pueden agruparse en seis tipos principales (MOLESTINA, 1988; UNESCO, 1951) según se presenta en la Tabla 3:

Tipos de escritos científicos	Definición y características
<p>ENSAYO</p>	<p>Escrito basado en un problema científico o en un grupo de problemas de magnitud considerable. El propósito es tratar un problema mayor tan definitivamente como sea posible. A menudo son evidentes las amplias interrelaciones de muchas ciencias. La presentación varía con la materia, pero, en un buen número de casos, el énfasis es en la teoría.</p>

ARTÍCULO	Escrito basado en una sola investigación. El propósito es contribuir al progreso de la ciencia o tecnología. Está redactado en tal forma que un investigador competente, basándose exclusivamente en las indicaciones que figuran en ese texto, pueda: 1) reproducir los experimentos y obtener los resultados que se describen con errores iguales o inferiores al límite superior indicado por el autor; 2) repetir las observaciones y juzgar las conclusiones del autor; y 3) verificar la exactitud de los análisis y deducciones que han permitido al autor llegar a sus conclusiones.
NOTA TÉCNICA	Escrito que proporciona información de resultados preliminares o de investigaciones en marcha. Si bien aporta información científica nueva, su redacción no permite a sus lectores verificar esa información en las condiciones indicada para el artículo. Corresponde a lo que la UNESCO llama “publicaciones provisionales” o “notas iniciales”.
REVISIÓN DE LITERATURA	Escrito basado en un análisis de lo publicado sobre un problema dado. El propósito es definir el estado actual de ese problema y evaluar la investigación hecha hasta el momento de escribirlo. Está presentado en términos de las fases del problema; avances hechos por investigaciones individuales o en grupos; cambios en la teoría o nuevas luces sobre ella; contradicciones sin resolver, enigmas, etc.; y direcciones y tendencias futuras. Corresponde a lo que la UNESCO llama “estudios recapitulativos”. Los libros son por lo general revisiones amplias de literatura.
INFORME	Escrito basado en la “necesidad de saber” de un cliente, superior o grupo directivo. Generalmente es más una herramienta de administración que una contribución científica. Está presentado, usualmente, en términos del progreso exacto realizado (con énfasis mínimo en cómo fue hecho el trabajo); el significado del progreso; etapas siguientes en la experimentación con énfasis en cómo se debe manejar la próxima etapa.
RESEÑA DE LIBROS	Escrito basado sobre un conocimiento especializado del campo sobre el que trata el libro. El tipo analítico de revisión tiene un tono judicial y busca evaluar los méritos de un libro en lo que respecta a su seriedad científica, los valores específicos que ofrece, el grado con que el libro alcanza sus objetivos y su rango de importancia en el área de estudio al que pertenece.

**Tabla 3: Tipos de textos científicos. Fuente: MOLESTINA et al., 1988; UNESCO, 1951.**

Una vez descritos los anteriores conceptos, se sabe cuáles son las características que debería tener un texto científico en general, lo cual es necesario para poder efectuar su producción. Por otro lado, se debería prestar especial atención a los tipos de escritos científicos anteriormente nombrados, ya que la solución a este problema se centrará en dichos tipos de escritos.

### 3.2 Conceptos relacionados a la propuesta de solución

Para poder entender mejor la solución propuesta, es necesario tener conocimiento sobre el área de la ciencia informática a la que pertenece la tecnología propuesta. A continuación, se definirán los conceptos relacionados a dicha área.

#### ➤ Procesamiento del Lenguaje Natural (PLN)

El procesamiento de lenguaje natural (PLN) consiste en el uso de computadoras para entender lenguajes (naturales) humanos tales como español, inglés, francés o japonés. Por *entender* no se quiere decir que el computador tenga pensamientos, sentimientos y conocimientos humanizados, sino que el computador pueda reconocer y usar información expresada en lenguaje humano (COVINGTON, 1994).



Un sistema de PLN es aquel que encapsula un modelo del lenguaje natural en algoritmos apropiados y eficientes, en donde las técnicas de modelado están ampliamente relacionadas con eventos en muchos otros campos, incluyendo (MANARIS & SLATOR, 1996):

- ✓ Ciencia de la computación, la cual provee métodos para representar modelos, diseñar e implementar algoritmos para herramientas de software.
- ✓ Lingüística, la cual contribuye con nuevos modelos lingüísticos y procesos.
- ✓ Matemática, la cual identifica modelos formales y métodos.
- ✓ Neurociencia, la cual explora los mecanismos mentales y otro tipo de actividades físicas.

Entre estos campos, la lingüística ha aportado el conocimiento de las lenguas naturales. Este conocimiento dentro de un sistema de PLN puede ser dividido en niveles definidos en términos de la característica declarativa (qué) y procedural (cómo), tal como se muestra en la Tabla 4 (MANARIS & SLATOR).

Nivel	Características	
	Declarativo (qué)	Procedural (cómo)
Fonológico	Sonidos hablados	Formar morfemas
Morfológico	Unidades de palabras, Palabras	Formar palabras, Derivar unidades de significado
Sintáctico	Roles estructurales de palabras (o colección de palabras)	Formar oraciones
Semántico	Significado independiente del contexto	Derivar significado de oraciones
Discurso	Roles estructurales de oraciones (o colección de oraciones)	Formar diálogos
Pragmático	Significado dependiente del contexto	Derivar significado de oraciones relativo al discurso redundante

**Tabla 4: Niveles de conocimiento en el PLN.**

Este conocimiento lingüístico se ha incorporado a los sistemas PLN desde los años 60 y, actualmente, se ha convertido en uno de los componentes más importantes de estos sistemas. Debido a esto, se ha definido un área del conocimiento llamado Lingüística Computacional, apoyado por la Asociación para la Lingüística Computacional.

Como nota adicional, sería importante saber los diversos problemas que implica el uso de los sistemas PLN (MANARIS & SLATOR, 1996).

- ✓ **Inexactitud**, incluyendo errores ortográficos, signos de puntuación incorrectos, palabras transpuestas, y oraciones agramaticales.
- ✓ **Incompletitud**, incluyendo construcciones elípticas, anáforas, etc.
- ✓ **Imprecisión**, incluyendo el uso de términos relativos sin un punto específico de referencia y el uso de términos cualitativos
- ✓ **Ambigüedad**, debido a que pueden surgir múltiples interpretaciones en cualquier nivel del conocimiento lingüístico (ver Tabla 4). La ambigüedad puede ser resuelta usando el conocimiento de un nivel más alto.

### ➤ **Lingüística Computacional**

La Lingüística Computacional es una disciplina que trata básicamente de dos cosas: lenguas naturales y computadoras. Muchas líneas de investigación comparten ambos objetivos aunque desde perspectivas diferentes. Como siempre hay que enfrentarse con el objeto de estudio y con la delimitación de las terminologías de las ciencias, hay que dejar claro que la lingüística computacional es equivalente al PLN (MORENO, 1998).

Tanto el PLN como la lingüística computacional tratan del desarrollo de programas de ordenador que simulan la capacidad lingüística humana. La lingüística computacional es el estudio de los sistemas de computación utilizados para la comprensión y la generación de las lenguas naturales (GRISHMAN, 1986); mientras que el objetivo de la investigación del PLN es crear modelos computacionales del lenguaje lo suficientemente detallados que permitan escribir programas informáticos que realicen las diferentes tareas en donde interviene el lenguaje natural (ALLEN, 1995). Como se puede apreciar, PLN y lingüística computacional son equivalentes. Sin embargo, en este trabajo se utilizará más el término PLN, pues suele ser mejor entendido.

Por otro lado, es importante conocer las principales aplicaciones prácticas de la lingüística computacional, tales como (MORENO, 1998):

- ✓ **Sistemas que tratan de emular la capacidad humana de procesar lenguas naturales:** Dentro de este grupo las más importantes son: Traducción automática, Recuperación y extracción de Información, Interfaces hombre-máquina.
- ✓ **Sistemas que ayudan en las tareas lingüísticas:** Este grupo está formado por herramientas que pueden ser utilizadas por los lingüistas para facilitarles ciertas tareas complejas. Algunas aplicaciones de este tipo son: Herramientas de análisis textual, Herramientas de manejo de corpus, Bases de datos lexicográficas.
- ✓ **Programas de ayuda a la escritura y composición textual:** Las aplicaciones comprendidas en este grupo han sido ampliamente desarrolladas y cualquier usuario habitual de un procesador de texto está familiarizado con ellas: Correctores ortográficos, Correctores sintácticos y de estilo.
- ✓ **Enseñanza asistida por computador:** Este es un campo de aplicación en continua expansión y que tiene varias vertientes. La más importante es la de los programas educativos para la enseñanza de las lenguas extranjeras.

Una vez descritos los anteriores conceptos, se tiene una idea más clara de la herramienta que se utilizará como solución la redacción de resúmenes de textos científicos. Cabe resaltar que la herramienta informática que se propone hacer en este proyecto estaría considerada dentro del conjunto de sistemas que ayudan en las tareas lingüísticas.

### ➤ **Aprendizaje de máquinas**

Para entender mejor el camino que se seguirá para el diseño e implementación de la herramienta, es necesario conocer sobre el aprendizaje automático de máquinas.

El principal objetivo del “aprendizaje automático” consiste en otorgar a la máquina la habilidad de mejorar gradual y paulatinamente su comportamiento sobre la base de nueva información recibida. Se trata de conseguir que los sistemas puedan deducir conceptos o información que no se le hubiera dado explícitamente, que aprendan de sus errores y que corrijan sus equivocaciones, beneficiándose de su propia experiencia. En el caso ideal, lo único que se tendría que codificar inicial y explícitamente sería una amplia “base de conocimientos generales” y algunas rutinas o reglas heurísticas que le sirvan de ayuda para adquirir, por su propia cuenta, nueva información externa y comprender nuevas situaciones, a medida que la máquina vaya adquiriendo nuevas “experiencias” (MORIELLO, 2004).

Según (MORIELLO, 2004), hay tres tipos de aprendizaje de máquina. El más elemental para una computadora es el “aprendizaje por implantación”, en donde el conocimiento se absorbe por pura “memorización”, así la máquina puede adquirir conocimiento con mucha mayor facilidad y rapidez que el ser humano e incorporarlo, sin riesgo al olvido, a su memoria. Un poco más complicado es el “aprendizaje por deducción”, en el cual se parte de reglas generales y se llega a la determinación de hechos específicos o al perfeccionamiento de estrategias ya desarrolladas. Por último, se encuentra el “aprendizaje por inducción”, en el cual se le suministran datos a la computadora y ésta extrae el conocimiento a partir de ellos.

En el último método “aprendizaje por inducción”, se verifican dos modalidades: a partir de ejemplos (supervisados) y a partir de observaciones (no supervisado). En el primer caso se presentan ejemplos correctos e incorrectos para que la máquina por sí misma extraiga reglas, leyes o conceptos de más alto nivel. En el segundo caso, la máquina solo “observa”, debiendo descubrir automáticamente los rasgos comunes en el grupo de datos observados, a fin de poder clasificarlos (MORIELLO, 2004).

### ➤ **Aprendizaje supervisado vs. Aprendizaje no supervisado**

En esta sección se mencionará con un poco más de detalle la diferencia entre ambas modalidades de aprendizaje automático (“supervisado” y “no supervisado”). La distinción entre ambos se ve en la forma cómo la máquina aprendiz clasifica los datos.

En los algoritmos supervisados las clases están predeterminadas. Estas clases pueden ser concebidas como un conjunto finito de elementos, previamente creado por un humano. En la práctica, un cierto segmento de datos será “etiquetado” con estas clases. La tarea de la máquina aprendiz es buscar por patrones y construir modelos matemáticos. Estos modelos luego se evalúan sobre la base de su capacidad predictiva en relación a las medidas de la varianza en los datos (MONK, s/f). Algunos ejemplos de técnicas de aprendizaje supervisado son los siguientes (MANNING, 2000):

- Árboles de decisiones
- Clasificadores de máxima entropía
- Clasificadores bayesianos ingenuos

Por otro lado, los algoritmos no supervisados no disponen de clases predeterminadas. De hecho, la tarea básica de aprendizaje no supervisado es

desarrollar “etiquetas de clases” de forma automática. Los algoritmos no supervisados buscan similitud entre las piezas de los datos con el fin de determinar si pueden ser caracterizados como la formación de un grupo. Estos grupos se denominan “clusters”, y hay toda una familia de técnicas de “clustering” para lograr el aprendizaje de máquina (MONK, s/f).

### ➤ Clasificación supervisada

La clasificación es la tarea de elegir la etiqueta de clase correcta para una entrada (input) dada. En las tareas de clasificación básicas, cada entrada se encuentra aislada de todas las demás entradas, y el conjunto de etiquetas se define por adelantado. Algunos ejemplos de las tareas de clasificación son:

- Decidir si un correo es spam o no.
- Decidir cuál es el tema de un artículo de prensa, a partir de una lista fija de áreas temáticas tales como "deportes", "tecnología" y "política".
- Decidir si un hecho determinado de la palabra “banco” se utiliza para referirse a la orilla de un río, una institución financiera, el acto de inclinar hacia un lado, o el acto de depositar algo en una institución financiera.

Un clasificador se llama supervisado si se construye sobre la base de un “corpus<sup>1</sup> de entrenamiento” que contiene la etiqueta correcta para cada entrada. El marco utilizado por clasificación supervisada se muestra en la Figura 1.

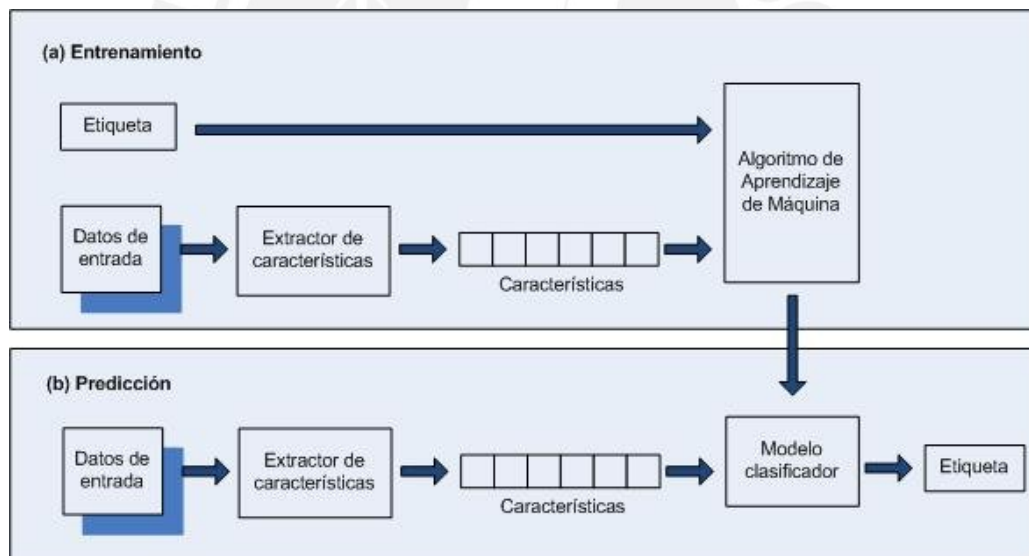


Figura 1: Marco de clasificación supervisada. Basado en un gráfico de NLTK, s/f.

Como se puede observar en la **Figura 1**, existen 2 fases en la clasificación supervisada:

- Durante el entrenamiento, un extractor de características se utiliza para convertir cada valor de entrada en un conjunto de características. Estos conjuntos de características capturan la información básica acerca de cada entrada que se debe utilizar para clasificarlo. Los pares de conjuntos de

<sup>1</sup> Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación (RAE, 2012).

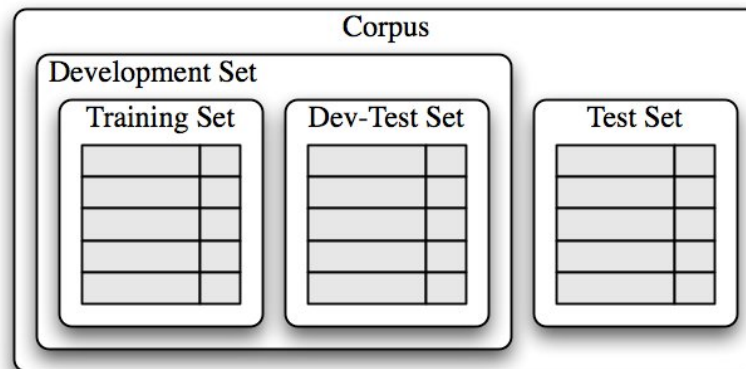
características y etiquetas se introducen en el algoritmo de aprendizaje de máquina para generar un modelo.

- b) Durante la predicción, el mismo extractor de características se utiliza para convertir las entradas “no previstas” en conjuntos de características. Estos conjuntos de características (sin etiquetas asignadas) se incorporan después en el modelo. Por último, el modelo genera las etiquetas (correctas) para dichos conjuntos de características.

La selección de características relevantes y decisión de cómo codificarlos por un método de aprendizaje puede tener un enorme impacto en la capacidad del método de aprendizaje para extraer un buen modelo. Gran parte del trabajo interesante en la construcción de un clasificador es decidir qué características podrían ser relevantes y cómo podemos representarlas. Aunque a menudo es posible obtener un rendimiento decente con un conjunto bastante simple y obvio de características, hay usualmente beneficios significativos por usar cuidadosamente características construidas sobre la base de un conocimiento profundo de la tarea en cuestión (NLTK, s/f).

Una vez que un conjunto inicial de características se ha escogido, un método muy productivo para perfeccionar el conjunto de características es el análisis de errores.

En primer lugar, se selecciona un Development Set, el cual se subdivide en Training Set, que se usa para entrenar el modelo, y Dev-Test Set, que se utiliza para llevar a cabo el análisis de errores. Por último, se selecciona un Test Set, que se utiliza para la evaluación final del sistema. Es una buena práctica que se cuente con un Dev-Test independiente para el análisis de errores, en lugar de utilizar el Test Set. La división de los datos del corpus en diferentes subconjuntos se muestra en la **Figura 2**.



**Figura 2: Organización del corpus de datos para el entrenamiento de clasificadores supervisados. Fuente: NLTK, s/f.**

### ➤ Evaluación del modelo de aprendizaje de máquina

Para decidir si un modelo de clasificación es preciso capturando un patrón, se debe evaluar dicho modelo. El resultado de esta evaluación es importante para decidir qué tan confiable es el modelo y con qué propósito lo podemos usar. La evaluación también puede ser una herramienta eficaz para guiarnos en la toma de futuras mejoras en el modelo.

La métrica más simple que se puede utilizar para evaluar un clasificador, la precisión, mide el porcentaje de entradas en el Test Set que el clasificador etiquetó correctamente (NLTK, s/f). Por ejemplo, un clasificador de género de nombres que predice el nombre correcto 60 veces en un Test Set, que contiene 80 nombres, tendría una precisión de  $60/80 = 75\%$ .

Cuando se interpreta el puntaje de precisión de un clasificador, es importante tener en consideración las frecuencias de las etiquetas de clase individuales en el Test Set, ya que se podría dar el caso de que el puntaje de precisión obtenido no sea exacto, y por ende, genere una mala interpretación de resultados.

Otro caso donde el puntaje de precisión puede ser engañoso es en las tareas de "búsqueda", como la recuperación de la información, donde estamos tratando de encontrar los documentos que son relevantes para una tarea en particular. Dado que el número de documentos irrelevantes es mucho mayor que el número de los documentos pertinentes, la puntuación de precisión para un modelo que etiqueta cada documento como irrelevante sería muy cerca de 100%.

Para evitar los casos mencionados anteriormente, es necesario emplear un conjunto diferente de medidas. En la **Tabla 5**, se muestra una tabla base para comprender las definiciones necesarias para obtener métricas, tales como Precision y Recall.

Clase C		Manual	
		SÍ	NO
Automática	SÍ	TP	FP
	NO	FN	TN

**Tabla 5: Verdaderos y Falsos Positivos y Negativos.**

Teniendo como referencia la **Tabla 5**, se procederá a explicar cada una de éstas cuatro categorías.

- Verdaderos positivos (TP), para la clase C, son instancias pertenecientes a la clase C, que se clasifican correctamente en la clase C.
- Verdaderos negativos (TN), para la clase C, son instancias no pertenecientes a la clase C, y que no se clasifican como clase C.
- Falsos positivos (FP), para la clase C, son instancias no pertenecientes a la clase C, pero que se clasifican como clase C.
- Falsos negativos (FN), para la clase C, son instancias pertenecientes a la clase C, pero que no se clasifican como clase C.

Teniendo en cuenta estos cuatro números, se puede definir las siguientes métricas:

- Precisión (P), para la clase C, es un valor entre 0 y 1. Su valor aumenta cuando hay pocos falsos positivos. Mide que las instancias clasificadas como clase C sean realmente de la clase C, aunque haya instancias de la clase C que se clasifiquen como otra clase. Su fórmula es  $TP/(TP+FP)$ .

- **Recall (R)**, para la clase C, es un valor entre 0 y 1. Su valor aumenta cuando hay pocos falsos negativos. Mide que las instancias de la clase C se clasifiquen como clase C, aunque otras instancias también se clasifiquen como clase C sin serlo. Su fórmula es  $TP/(TP+FN)$ .
- **F-Measure (o F-Score)** combina Precision y Recall para obtener un puntaje único. Éste es definido como la media armónica de Precision y Recall. Su fórmula es  $(P \times R)/(P+R)$ .

## 4 Estado del arte

En esta sección, se presentará diversas soluciones al problema planteado por medio de productos comerciales o investigaciones realizadas.

### 4.1 Formas exactas de resolver el problema

El método tradicional para resolver el problema de escritura de textos científicos es el de acudir a libros de redacción, ya que en éstos se muestra el proceso que el estudiante debe seguir para redactar de forma adecuada un texto académico o científico.

En el libro de Redacción y Comunicación de la PUCP (Equipo de trabajo del curso de Redacción y Comunicación de EE.GG.CC de la PUCP, 2008) se señala que en un texto académico, tanto la organización de las ideas en párrafos como su adecuada disposición son resultado del cumplimiento de “pasos previos” que nos facilitan la redacción de un texto coherente. Los pasos son los siguientes:

#### Primero: Tratamiento de la información

- ✓ Obtener información sobre el tema que se quiere desarrollar.
- ✓ Delimitar el tema

#### Segundo: Organización de las ideas en un esquema

- ✓ Seleccionar las ideas pertinentes según el tema a tratar.
- ✓ Establecer la división del texto en párrafos.
- ✓ Asociar una idea principal a cada párrafo.
- ✓ Subordinar cada idea secundaria al párrafo que le corresponde.

#### Tercero: Redacción y revisión de la versión preliminar

- ✓ Redactar el texto siguiendo la estructura previamente propuesta en el esquema.
- ✓ Utilizar adecuadamente referencias y conectores.
- ✓ Revisar el uso adecuado de la ortografía y la variedad formal del texto.

#### Cuarto: Redacción de la versión final del texto

- ✓ Revisar la consistencia de las definiciones, generalizaciones, ejemplos, relaciones de causalidad y clasificaciones.
- ✓ Revisar el óptimo empleo de conectores y referentes poniendo atención a las posibles redundancias o incoherencias.

También es importante tener en cuenta los siguientes aspectos:

- ✓ El tema central del texto debe identificarse en forma fácil y clara.
- ✓ Cada párrafo debe desarrollar una idea principal.
- ✓ La conexión entre las ideas (ilación) debe ser explícita para el lector.

- ✓ El texto debe satisfacer los requisitos de la variedad formal: buena ortografía, puntuación, léxico académico y adecuada construcción de enunciados.

Como se puede observar, los pasos anteriormente mencionados son para elaborar textos académicos en general. Es necesario tener estos pasos en cuenta para tener una idea general de cómo redactar textos científicos.

Para lograr una buena redacción de este tipo de textos, también es necesario conocer su estructura. A continuación, se mostrará una estructura ideal para los resúmenes de textos científicos, así como los pasos pertinentes que se deben seguir para formar los diversos componentes de la estructura.

Componentes de la estructura	Pasos
Contexto	<ol style="list-style-type: none"> <li>1. Declarar la prominencia del tema</li> <li>2. Familiarizar términos, objetos y procesos</li> <li>3. Citar resultados de investigaciones anteriores</li> <li>4. Presentar hipótesis</li> </ol>
Brecha	<ol style="list-style-type: none"> <li>1. Citar problemas/dificultades</li> <li>2. Citar necesidades/requisitos</li> <li>3. Citar la ausencia o falta de investigación anterior</li> </ol>
Propósito	<ol style="list-style-type: none"> <li>1. Presentar el propósito principal</li> <li>2. Detallar/Especificar el propósito</li> <li>3. Presentar más propósitos</li> <li>4. Presentar el propósito con la metodología</li> <li>5. Presentar el propósito con los resultados</li> </ol>
Metodología	<ol style="list-style-type: none"> <li>1. Listar criterios o condiciones</li> <li>2. Citar/Describir materiales y métodos</li> <li>3. Justificar la elección de materiales y métodos</li> </ol>
Resultado	<ol style="list-style-type: none"> <li>1. Describir los resultados</li> <li>2. Indicar los resultados</li> <li>3. Comentar/Discutir los resultados</li> </ol>
Conclusión	<ol style="list-style-type: none"> <li>1. Presentar conclusiones</li> <li>2. Presentar contribuciones/valor de investigación</li> <li>3. Presentar recomendaciones</li> <li>4. Presentar lista de tópicos abordados en el trabajo</li> </ol>

**Tabla 6: Estructura ideal de un resumen de texto científico. Fuente: SciPo, 2000.**

Tener conocimiento de la estructura del resumen de un texto científico, y de los pasos a seguir para lograr redactar bien un texto académico en general, ayudará al estudiante universitario a aprender a redactar textos científicos de calidad.

#### 4.2 Productos comerciales para resolver el problema

En esta sub-sección, se presentará diversos software comerciales que solucionan el problema planteado, tanto de manera parcial como total. Se podrán apreciar sus diversas características, así como los problemas específicos que solucionan.



### ➤ **E-rater & Criterion**

E-rater (Electronic Essay Rater), desarrollado por The Educational Testing Service (ETS), evalúa la calidad de un ensayo teniendo como base las características lingüísticas en el texto (RUDNER et al., 2000). Este programa usa técnicas del Procesamiento de Lenguaje Natural (PLN), las cuales identifican señales léxicas y sintácticas específicas en un texto para analizar el ensayo (BURSTEIN, 2003). E-rater necesita “entrenarse” con alrededor de 450 textos pre-evaluados para poder analizar nuevos textos (ATTALI & BURSTEIN, 2006).

Por otro lado, Criterion (Online Essay Evaluation Service) es un sistema web cuyo fin es evaluar las habilidades relacionadas a la escritura de los usuarios, también les proporciona una puntuación de acuerdo a su calidad de escritura, y hace la retroalimentación del texto (RUDNER et al., 2000). Criterion se basa en el programa E-rater para puntuar los textos; además, contiene una aplicación llamada Critique que incluye un grupo de programas que identifica los errores gramaticales, los usos del lenguaje, la estructuración del discurso y la ineficiencias en los estilos de los textos (ATTALI & BURSTEIN, 2006).

### ➤ **IntelliMetric & MY Access!**

IntelliMetric es un sistema AES (Automated Essay Scoring) conocido como la primera herramienta de puntuación de ensayos basada en Inteligencia Artificial (IA) (ELLIOT, 2003). Al igual que E-rate, este sistema utiliza técnicas de PLN.

IntelliMetric evalúa alrededor de 300 de características relacionadas a la semántica, sintaxis y discurso en un ensayo, y necesita ser entrenado con 300 ensayos pre-evaluados antes de ser utilizado para el análisis de textos (ELLIOT, 2003).

Por otro lado, MY Access! es un sistema web de evaluación de la escritura. MY Access! está basado en el sistema AES IntelliMetric. El principal objetivo del programa es ofrecer a los estudiantes un entorno de escritura que proporciona una calificación inmediata y un diagnóstico de retroalimentación; en consecuencia, les permite revisar sus ensayos y los motiva a continuar escribiendo sobre el tema para mejorar su eficiencia en la escritura (RUDNER et al., 2000).

IntelliMetric y MY Access! fueron desarrollados por Vantage Learning, y usado por The Collage Board para fines educativos.

## **4.3 Productos no comerciales (de investigación) para resolver el problema**

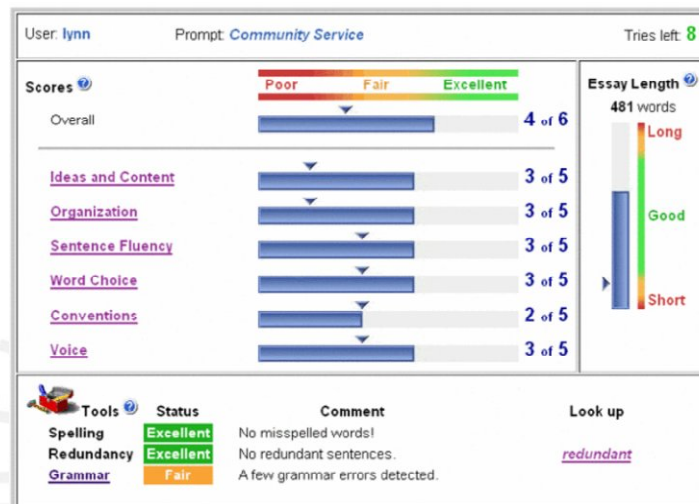
En esta sub-sección, se presentará diversos software no comerciales (de investigación) que solucionan el problema planteado, tanto de manera parcial como total. Se podrán apreciar sus diversas características, así como los problemas específicos que solucionan.

### ➤ **IEA (Intelligence Essay Assessor)**

IEA analiza y anota un texto mediante un método de análisis semántico llamado Análisis Semántico Latente (LSA – *Latent Semantic Analysis*). La ASL es definida como “un modelo estadístico del uso de palabras que permite comparaciones de semejanza semántica entre partes de la información textual” (FOLTZ, 1996).

A diferencia de otros sistemas AES, IEA se enfoca más en las características relacionadas con el contenido que aquellas relacionadas con la forma. El sistema usa un enfoque basado en LSA para evaluar principalmente la calidad del contenido de un ensayo; sin embargo, también incluye puntuación y retroalimentación en cuanto a la gramática, estilo y mecánicas. Cabe resaltar que este programa requiere alrededor de 100 textos pre-evaluados para poder analizar nuevos textos (LANDAUER et al., 2003).

En la **Figura 3**, se muestra un ejemplo de la función de retroalimentación que brinda IEA. Se puede ver que la aplicación brinda un puntaje de acuerdo a la calidad de escritura en cuanto a las ideas y contenido, organización, fluencia de la oración, elección de las palabras, entre otros.



**Figura 3: Sample Feedback Screen of IEA. Fuente: RUDNER et al., 2000.**

IEA fue creado por el psicólogo Thomas Landauer con la asistencia de Peter Foltz y Dareell Laham. IEA fue producida por Pearson Knowledge Analysis Technologies (RUDNER et al., 2000).

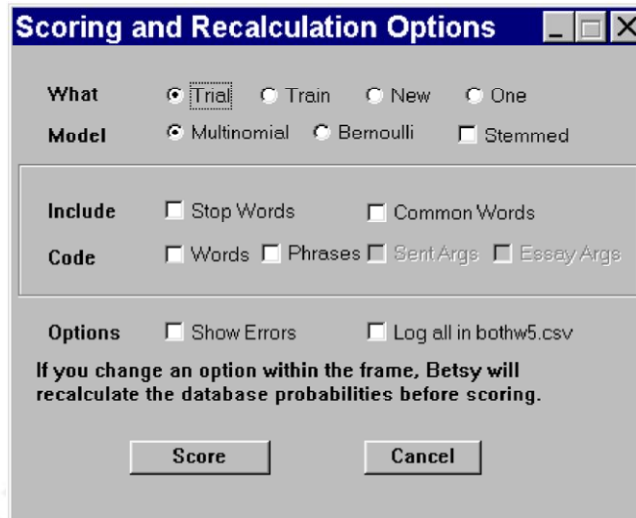
### ➤ **BETSY (Bayesian Essay Test Scoring sYstem)**

BETSY es uno de los pocos sistemas que es presentado como no comercial y está disponible para usar libremente. BETSY es un programa basado en ventanas que clasifica el texto basado en material pre-evaluado. Fue diseñado para que la puntuación de ensayos sea automática y puede ser aplicado a cualquier tarea de clasificación de texto (BETSY, s/f).

BETSY es una herramienta que puede ser empleada para evaluar composiciones cortas con una amplia gama de áreas conocimiento. Puede ser utilizada para obtener resultados de diagnóstico y puede ser adaptada para clasificar textos de múltiples habilidades (RUDNER & LIANG, 2002).

El problema de este programa es que requiere de al menos 1000 textos pre-evaluados para analizar nuevos textos (RUDNER et al., 2000). Esto genera desventajas con respecto a otros programas. Uno de ellos es que su entrenamiento es lento. Otro defecto es que se restringe al entorno de Windows.

En la **Figura 4**, se muestra una ventana de BETSY donde se configura los parámetros, tales como el modelo estadístico a utilizar, inclusión de palabras comunes, entre otros. Estos parámetros serán empleados para recalculer las probabilidades de la base de datos antes de realizar la calificación del texto ingresado por el estudiante.



**Figura 4: Scoring and recalculation window in BETSY. Fuente: RUDNER et al., 2000.**

BETSY fue desarrollado por el doctor Lawrence M. Rudner, con fondos otorgados por The U.S. Department of Education. Algunas evaluaciones de BETSY fueron posibles gracias a los fondos otorgados por Maryland State Department of Education.

#### ➤ SciPo

SciPo es una herramienta para redacción científica en portugués. Tiene como objetivo ayudar al estudiante en la escritura de resúmenes y presentaciones de textos académicos. Este sistema apoya en la estructuración de los textos según las directrices de la "buena escritura" propuestos por la literatura. Además, se puede consultar una base de datos que contiene ejemplos auténticos (y comentados) de introducciones, resúmenes de tesis y disertaciones en Ciencias de la Computación (SciPo, 2000).

SciPo tiene la capacidad de evaluar escritos de resúmenes e introducciones. En cada caso, el sistema establece una estructura ideal determinada para la evaluación del texto ingresado por el usuario. Un ejemplo de estructura se puede visualizar en la Tabla 6, la cual pertenece a los resúmenes de textos científicos.

En la **Figura 5**, se puede ver cómo es la interfaz con el que el usuario interactúa para la redacción y evaluación de sus resúmenes en portugués. Así también, se puede ver la estructura que debería tener un resumen, según SciPo. Finalmente, en la parte inferior, se puede observar que hay opciones donde el usuario puede solicitar críticas, sugerencia, así como ejemplos similares.

## Resumo - Seleção da estrutura

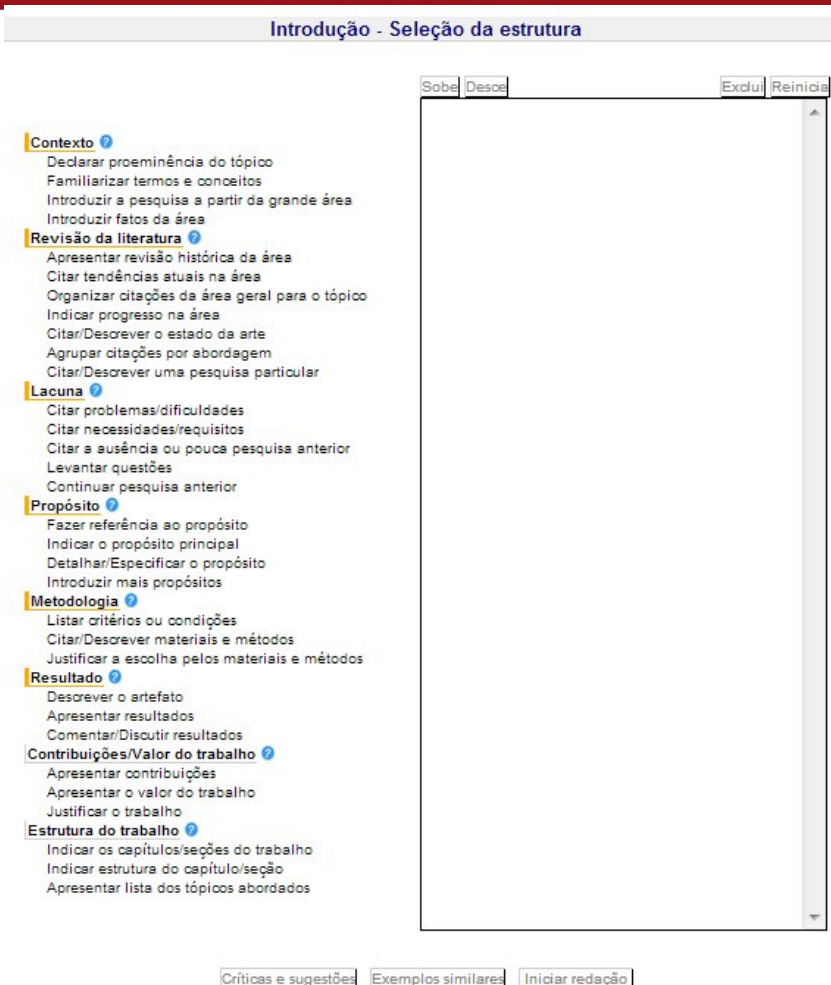
Sobe Desce Exclui Reinicia

- Contexto** ?
  - Declarar proeminência do tópico
  - Familiarizar termos e conceitos
  - Introduzir a pesquisa a partir da grande área
- Lacuna** ?
  - Citar problemas/dificuldades
  - Citar necessidades/requisitos
  - Citar a ausência ou pouca pesquisa anterior
- Propósito** ?
  - Indicar o propósito principal
  - Detalhar/Especificar o propósito
  - Introduzir mais propósitos
- Metodologia** ?
  - Listar critérios ou condições
  - Citar/Descrever materiais e métodos
  - Justificar a escolha pelos materiais e métodos
- Resultado** ?
  - Descrever o artefato
  - Apresentar resultados
  - Comentar/Discutir resultados
- Conclusão** ?
  - Apresentar conclusões
  - Apresentar contribuições/valor do trabalho
  - Apresentar recomendação

Críticas e sugestões Exemplos similares Iniciar redação

**Figura 5: Evaluación de resúmenes en portugués. Fuente: SciPo, 2000.**

Por otro lado, en la **Figura 6**, se puede ver cómo es la interfaz con el que el usuario interactúa para la redacción y evaluación de sus introducciones en portugués. Se puede observar que cuenta con las mismas funcionalidades del caso anterior, no obstante, establece una estructura diferente.



**Figura 6: Evaluación de introducciones en portugués. Fuente: SciPo, 2000.**

El sistema SciPo forma parte de la tesis doctoral de Valeria D. Feltrim, titulado “Apoyo computacional para la escritura científica en portugués”. Fue desarrollado en el Núcleo Interinstitucional de la Lingüística Computacional (NILC) (ICMC - USP / São Carlos), bajo la dirección del profesor. Dr. Maria das Graças Volpe Nunes (asesor) y el Prof. Dra. Sandra Maria Aluisio (co-director). Este trabajo fue apoyado por la FAPESP, CAPES y CNPq (SciPo, 2000).

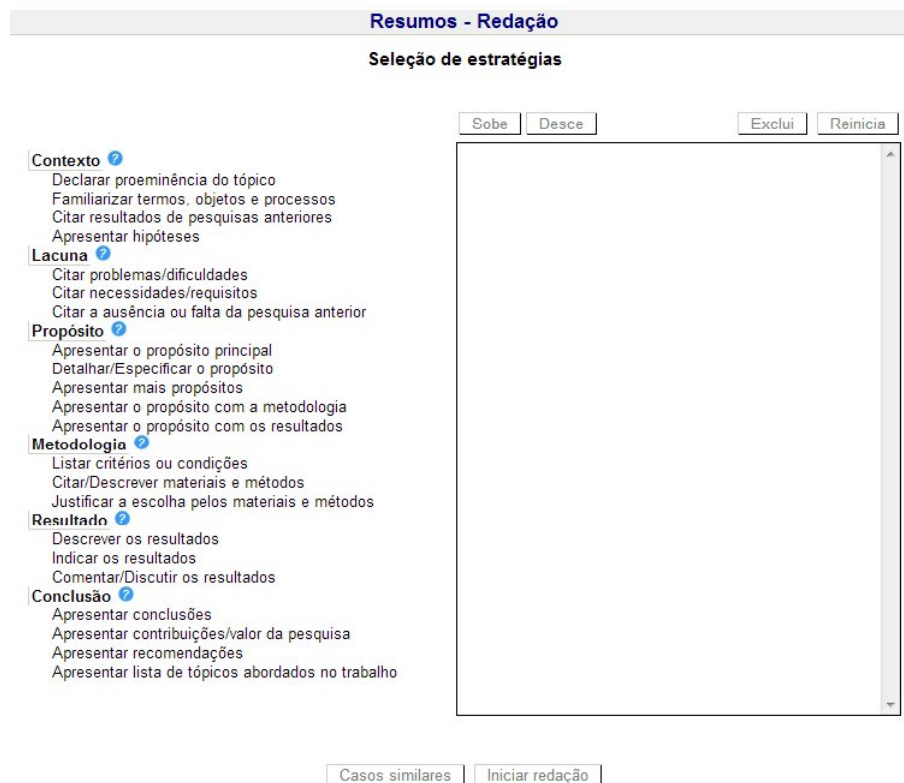
➤ **SciPo-Farmacía**

SciPo-Farmacía es una herramienta para redacción científica en inglés. Tiene como objetivo ayudar al estudiante en la escritura de resúmenes y presentaciones de textos académicos. Este sistema apoya en la estructuración de los textos según las directrices de la "buena escritura" propuestos por la literatura. Además, se puede consultar una base de datos que contiene ejemplos auténticos (y comentados) de resúmenes, presentaciones, métodos, resultados, discusiones y conclusiones de artículos en diversas sub-áreas de la ciencia (SciPo-Farmacía, s/f).

SciPo-Farmacía tiene la capacidad de evaluar escritos de resúmenes, introducciones, metodologías, resultados, discusiones y conclusiones. En cada

caso, el sistema establece una estructura ideal determinada para la evaluación del texto ingresado por el usuario.

A continuación, en la **Figura 7**, se puede ver cómo es la interfaz con el que el usuario interactúa para la redacción y evaluación de sus resúmenes en inglés. Además, se muestra la estructura que debería tener un resumen, según SciPo-Farmacía. Esta herramienta cuenta con la funcionalidad de brindar ejemplos similares si el usuario lo requiere.



**Figura 7: Evaluación de resúmenes en inglés. Fuente: SciPo-Farmacía, 2000.**

Como se había mencionado anteriormente, SciPo-Farmacía no sólo es capaz de evaluar resúmenes, sino también otros escritos. La diferencia entre uno y otro es la estructura del mismo; por otro lado, las funcionalidades desarrolladas, tales como la posibilidad de solicitar ejemplos similares, están presentes en todos los tipos de escrito.

El sistema SciPo-Farmacía se realizó en NILC. Este proyecto estuvo bajo la dirección de la profesora Sandra Maria Aluisio y el profesor Osvaldo Novais de Oliveira Jr., en colaboración con la Facultad de Ciencias Farmacéuticas de la USP, São Paulo, en particular con los profesores Adalberto Pessoa Jr. y Ana Campa. Un análisis textual de los artículos del instrumento se llevó a cabo por la lingüista Aline Maria Pacifico Manfrim, y posteriormente evaluado por el profesor Osvaldo Novais de Oliveira Jr. y la profesora Sandra Maria Aluisio (SciPo-Farmacía, s/f).

SciPo-Farmacía es un proyecto adaptado del sistema SciPo, tesis doctoral de Valeria D. Feltrim, titulado "Apoyo computacional para la escritura científica en portugués", desarrollado en el Núcleo Interinstitucional de la Lingüística

Computacional (NILC) (ICMC - USP / São Carlos), bajo la dirección del profesor. Dr. Maria das Graças Volpe Nunes (asesor) y el Prof. Dra. Sandra María Aluisio (co-director). En el caso de SciPo, se podrá evaluar la escritura de todos los componentes de un artículo científico (resúmenes, presentaciones, métodos, resultados, discusión y conclusiones) con el inglés como el idioma de destino (SciPo-Farmacia, s/f).

#### 4.4 Conclusiones sobre el estado del arte

En lo acápite anteriores se pudo ver los distintos tipos de solución, uno de ellos utilizando el método tradicional y otros mediante el uso de software.

Históricamente, las personas han usado el método tradicional para mejorar su redacción de textos científicos, pues los software dedicados a la evaluación de la escritura son tecnologías que han brotado recientemente. Actualmente, a pesar de que se cuente con software de evaluación de escritura, las personas continúan optando por el método tradicional debido a su inexperiencia en el área informática. Cabe resaltar que la falta de experiencia no es la única barrera de aprendizaje de la redacción mediante software, sino también el “idioma”.

La mayoría de software dedicados a la evaluación de la escritura, son desarrollados en inglés, portugués, ruso, etc., y funcionan de manera ideal para esos idiomas. Son muy pocos los software que están especializados en el estudio de la lengua española, incluso no hay mucha información de software (en español) que sean capaces de evaluar la estructura de un texto científico en español.

En cuanto a los software mencionados anteriormente, se realizó la siguiente tabla:

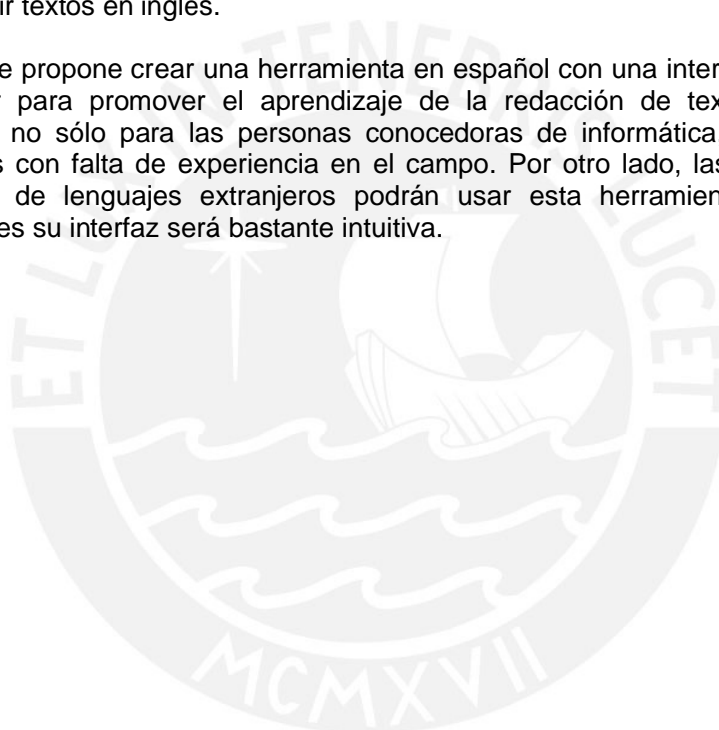
Sistema	Desarrollador	Técnica	Enfoque	Aplicación instructiva	N° de ensayos requeridos para su entrenamiento
IEA	Landauer, Laham & Foltz	Análisis Semántico Latente (ASL)	Contenido	N/A	100
E-rater	ETS development team	Procesamiento de lenguaje natural (PLN)	Estilo y contenido	Criterion	400-500
IntelliMetric	Vantage Learning	Procesamiento de lenguaje natural (PLN)	Estilo y contenido	MY Access!	300
BETSY	Rudner, L.	Clasificación bayesiana	Estilo y contenido	N/A	1000
SciPo	Núcleo Interinstitucional de la Lingüística Computacional	Procesamiento de lenguaje natural (PLN)	Estilo y contenido	SciPo	No especificado
SciPo-Farmacia	Núcleo Interinstitucional de la Lingüística Computacional	Procesamiento de lenguaje natural (PLN)	Estilo y contenido	SciPo-Farmacia	No especificado

**Tabla 7: Características de los software dedicados a la evaluación de la escritura.**

Como se puede apreciar, existen herramientas que brindan apoyo a la redacción de textos. Estas herramientas no sólo se centran en la estructura del texto, sino también en temas relativos a la ortografía, gramática, sintaxis, etc. Sin embargo, la mayoría se encuentra en inglés, algunas en portugués. Una gran cantidad de personas en el Perú, al no conocer el idioma extranjero a la perfección, prefieren acudir a herramientas o soluciones que se encuentren en su idioma materno, pues de esta forma se evitan confusiones en cuanto a vocabulario (presencia de muchos términos técnicos), gramática (nivel avanzado), entre otros que podría dar el hecho de no conocer bien este idioma. Por ende, dejarían estos sistemas de lado.

Otro factor desfavorable para estos sistemas es que éstos usan documentos pre-evaluados en inglés u otro idioma diferente al español. Si una herramienta se entrena con textos de un idioma determinado, sólo podría corregir textos que se encuentren en dicho idioma. Es decir, una herramienta que se entrena con textos en inglés, solo podría corregir textos en inglés.

Es así, que se propone crear una herramienta en español con una interfaz amigable y fácil de usar para promover el aprendizaje de la redacción de textos científicos (resúmenes), no sólo para las personas conocedoras de informática, sino también para aquellas con falta de experiencia en el campo. Por otro lado, las personas sin conocimiento de lenguajes extranjeros podrán usar esta herramienta sin ningún problema, pues su interfaz será bastante intuitiva.





## CAPÍTULO 2

### 1 Objetivo general

Desarrollar una herramienta informática de apoyo a la escritura de resúmenes de textos científicos en español que utilice modelos de aprendizaje de máquina supervisado.

### 2 Objetivos específicos

- Objetivo 1: Formar un corpus de resúmenes de textos científicos en español que sirvan para el entrenamiento y prueba de los modelos de aprendizaje.
- Objetivo 2: Determinar las características (atributos) que serán extraídas de cada oración de los resúmenes de textos científicos para que sean utilizadas por los modelos de aprendizaje.
- Objetivo 3: Implementar un modelo clasificador y evaluar su desempeño utilizando métricas tales como Precision, Recall y F-Measure.
- Objetivo 4: Implementar una aplicación que clasifique automáticamente las oraciones de los resúmenes de textos científicos en español ingresados por el usuario, siguiendo una estructura pre-definida.

### 3 Resultados esperados

- Resultado 1 para el objetivo 1: Corpus de 44 resúmenes de textos científicos en español.
- Resultado 2 para el objetivo 2: Conjunto de 6 características (atributos) de las oraciones de los resúmenes para entrenar y probar los modelos de aprendizaje.
- Resultado 3 para el objetivo 3: Modelo clasificador con una precisión cercana al 65%.
- Resultado 4 para el objetivo 4: Aplicación que clasifica automáticamente componentes de la estructura de los resúmenes de textos científicos.

### 4 Herramientas, métodos y procedimientos

En esta sección, se mostrará un mapeo entre los resultados esperados y las herramientas que se usó a lo largo del proyecto. Posteriormente, se describirá de forma más detallada cada herramienta, así como también las metodologías que se usó para gestionar el proyecto y el desarrollo del software.

## 4.1 Mapeo

Resultados esperado	Herramientas, métodos y procedimientos a usarse
RE1: Corpus de resúmenes de textos científicos en español.	La <b>recopilación de textos e identificación de componentes de estructura</b> se basa en buscar resúmenes de textos científicos en español en repositorios de tesis, e identificar los componentes de su estructura según definida previamente a través de un manual de anotación.
RE2: Conjunto de características de los resúmenes para entrenar y probar los modelos de aprendizaje.	El paquete <b>FreeLing</b> consta de una librería que proporciona herramientas para análisis de textos en diferentes niveles lingüísticos y para diferentes idiomas, incluido el español.
RE3: Modelo clasificador.	<b>WEKA</b> es un software que contiene una colección de algoritmos de aprendizaje de máquina para las tareas de minería de datos.
RE4: Aplicación que clasifica automáticamente componentes de la estructura de los resúmenes de textos científicos.	<b>WEKA</b> es un software que contiene una colección de algoritmos de aprendizaje de máquina para las tareas de minería de datos.
	El paquete <b>FreeLing</b> consta de una librería que proporciona herramientas para análisis de textos en diferentes niveles lingüísticos y para diferentes idiomas, incluido el español.

**Tabla 8: Mapeo de resultados esperados y herramientas, métodos y procedimientos a usarse.**

## 4.2 Herramientas, métodos y procedimientos

### ➤ Recopilación de textos e identificación de componentes de estructura

En cuanto a la recopilación de textos, se eligió coleccionar específicamente resúmenes de textos científicos, tales como tesis y disertaciones de las diferentes áreas de la Informática, las cuales cuentan con una estructura bien definida, a comparación de otras como son el caso de textos relacionados al área de humanidades siguiendo una metodología similar a la aplicada en (FELTRIM et al., 2001). Se decidió este tipo de textos no sólo por su estructura bien definida, sino también por la facilidad de recopilación.

Por otro lado, se sabe que la estructura esquemática de textos científicos se ha venido investigando durante un tiempo. Hay varios artículos que tratan de la escritura científica en líneas generales, mientras que otros estudios muestran la estructura esquemática de textos más detalladamente (FELTRIM et al., 2001). A pesar de que estos últimos sean estudios de textos en inglés y portugués, se ha constatado que los componentes de estructura encontrados también aplican para el español.

En cuanto al RE1, como se había mencionado anteriormente, se recopiló resúmenes de textos científicos. Así también, se definirán los componentes de acuerdo a lo encontrado en los estudios anteriormente mencionados. La estructura óptima será conformada por tres componentes principales (propósito, metodología y resultado), y tres componentes opcionales (contexto, brecha y conclusión). Éstos fueron identificados manualmente en cada texto del corpus, previamente recopilado.

## ➤ WEKA

**WEKA** es un software que contiene una colección de algoritmos de aprendizaje de máquina para las tareas de minería de datos. Este programa es de libre distribución y difusión. Además, ya que WEKA está programado en Java, es independiente de la arquitectura, ya que funciona en cualquier plataforma sobre la que haya una máquina virtual Java disponible (WEKA 3, s/f).

Para el RE3, se usó una funcionalidad de WEKA llamada “Classify”, cuyo objetivo es clasificar por varios métodos los datos ya cargados (WEKA 3, s/f). En la **Figura 8**, se puede apreciar que se definirá un clasificador, el cual puede ser configurado a nuestro gusto, ya que podemos elegir los filtros y los argumentos con los que se ejecutará. El resultado de la ejecución se puede apreciar en la sección “Classifier Output”.

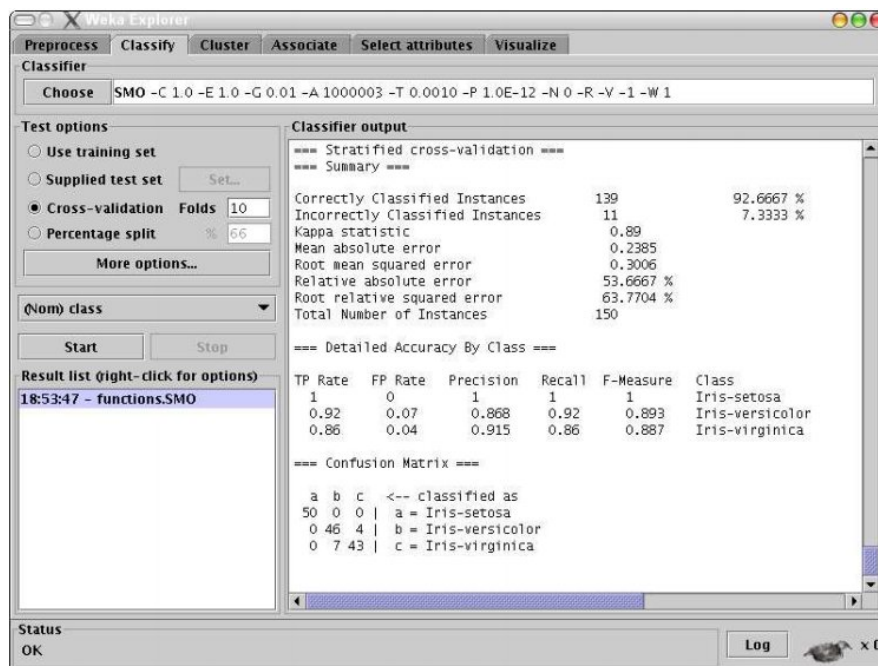


Figura 8: El modo de Clasificación. Fuente: WEKA 3, s/f.

## ➤ FreeLing

**FreeLing** es una librería de código abierto para el procesamiento multilingüe, que proporciona una amplia gama de funcionalidades de análisis para varios idiomas (FREELING, s/f). El proyecto se estructura como una librería que puede ser llamada desde cualquier aplicación de usuario que requiera servicios de análisis del lenguaje (PADRÓ, 2011). En la **Tabla 9**, se muestra una lista de servicios de análisis lingüísticos disponibles para cada idioma.

	Asturi ano	Catal án	Galé s	Castell ano	Ingl és	Galle go	Italia no	Portug ués	Ru so
Tokenización	X	X	X	X	X	X	X	X	X
Divisor de oraciones	X	X	X	X	X	X	X	X	X
Detección de número		X		X	X	X	X	X	X
Detección de fecha		X		X	X	X		X	X
Diccionario morfológico	X	X	X	X	X	X	X	X	X
Reglas de afijos	X	X	X	X	X	X	X	X	
Detección de múltiples palabras	X	X	X	X	X	X	X	X	
Detección de nombres básicos de entidades	X	X	X	X	X	X	X	X	X
Detección de nombres B-I-O de entidades				X	X	X			
Clasificación de nombres de entidades				X	X				
Detección de cantidad		X		X	X	X		X	X
Etiquetado PoS	X	X	X	X	X	X		X	X
Sentido de anotación WN		X		X	X				
Sentido de desambiguación UKB		X		X	X				
Análisis superficial	X	X		X	X	X		X	
Análisis completo/dependiente	X	X		X	X	X			
Resolución de correferencia					X				

**Tabla 9: Servicios de análisis lingüísticos disponibles para cada lengua. Fuente: PADRÓ & STANILOVSKY, 2012.**

En cuanto al RE2, esta librería facilitará el análisis lingüístico de cada componente de la estructura de los resúmenes de los textos científicos previamente revisados. Además, ayudará en la identificación de estos componentes en los textos nuevos, pertenecientes al conjunto de testeo.

### 4.3 Metodologías

#### ➤ Scrum

Scrum es un marco de trabajo de procesos que ha sido utilizado para gestionar el desarrollo de productos complejos desde principios de los años 90. Scrum no es un proceso o una técnica para construir productos; en lugar de eso, es un marco de trabajo dentro del cual se pueden emplear varios procesos y técnicas (SCRUM, 2011).

A continuación, se describirá algunos componentes de la metodología Scrum que serán utilizados para el proyecto de fin de carrera.

- Sprint  
El corazón de Scrum es el Sprint, un bloque de tiempo (time-box) de un mes o menos durante el cual se crea un incremento de producto “Hecho”, utilizable y potencialmente entregable. La duración de los Sprints es consistente a lo largo del esfuerzo de desarrollo. Cada nuevo Sprint comienza inmediatamente después de la finalización del Sprint previo (SCRUM, 2011). Se empleará el uso de Sprint en el proyecto de fin de carrera, pues se planea definir un lapso de tiempo no mayor a 3 semanas para la presentación de los entregables respectivos, cumpliendo lo acordado en los objetivos específicos.
- Cancelación del Sprint  
Un Sprint puede ser cancelado antes de que el bloque de tiempo llegue a su fin. En general, un Sprint debería cancelarse si no tuviese sentido seguir con él dadas las circunstancias (SCRUM, 2011). En el proyecto de fin de carrera, es probable que se dé este caso, ya que se puede presentar ciertos inconvenientes que imposibiliten la presentación del sprint. Así también, se puede presentar cambios radicales que hagan que el objetivo del Sprint quede obsoleto.
- Reunión de planificación de Sprint (Sprint Planning Meeting)  
El trabajo a realizar durante el Sprint es planificado en la Reunión de Planificación de Sprint. Este plan es creado mediante el trabajo colaborativo del Equipo Scrum al completo. La duración de la reunión es relativo a la duración del Sprint (SCRUM, 2011). El equipo estará conformado por el asesor y el tesista, y se tendrá una reunión con una duración de 1 hora a 2 horas. Se acordará qué funcionalidad del software será implementado en el siguiente Sprint, así como los documentos respectivos. También, se conversará acerca de cómo se conseguirá completar dicho Sprint.
- Revisión de Sprint (Sprint Review)  
Al final del Sprint se lleva a cabo una Revisión de Sprint, para inspeccionar el Incremento<sup>2</sup> del producto. Durante la Revisión de Sprint, el Equipo Scrum y los interesados colaboran acerca de lo que se ha hecho durante el Sprint. La duración de la reunión es relativo a la duración del Sprint (SCRUM, 2011). El equipo estará conformado por el asesor y el tesista, como se

---

<sup>2</sup> El Incremento es la suma de todos los elementos completados durante un Sprint y durante todos los Sprints previos.

mencionó anteriormente. Se tendrá una reunión con una duración de 1 hora, se hará en conjunto con la reunión de planificación de Sprint. Se revisará si lo acordado en el Sprint está completa o parcialmente hecho. Esto influirá en el acuerdo de entregables para el siguiente Sprint.

## 5 Alcance

El proyecto de fin de carrera se relaciona con la capacidad de redacción del sector universitario, específicamente la redacción de resúmenes de textos científicos. Se ha elegido este tema en especial debido a la gran importancia que tiene saber redactar en el campo social, académico y laboral en la actualidad.

Se trata de un proyecto de Procesamiento de Lenguaje Natural, rama de las Ciencias de la Computación. El proyecto se basa en la implementación de un software que ayude en la redacción de resúmenes de textos científicos en español. Se optó por enfocar el proyecto en los textos en español, no sólo por el dominio que se tiene sobre este idioma, sino también debido a que no existen muchas herramientas de ayuda en la redacción de estos textos.

En cuanto al aprendizaje de la herramienta, se tendrá un corpus de estudio que estará compuesto por una gama, no mayor a 50, de resúmenes de textos científicos en español. Cabe resaltar que éstos serán previamente revisados con el fin de garantizar su correcta redacción, y por ende un mejor funcionamiento de la herramienta en cuestión.

La herramienta será capaz de identificar los diversos componentes que conforman la estructura de los resúmenes de textos científicos ingresados por el usuario.

Con el desarrollo de esta herramienta, se busca que los universitarios puedan ver sus deficiencias en el tema y así mejorar su redacción de resúmenes de textos científicos de forma autodidacta y eficiente. Por otro lado, se desea, si es posible, poder llegar a ser un punto de partida para otros proyectos del mismo rubro con un alcance más profundo.

### 5.1 Limitaciones

- El sistema solo evaluará resúmenes de textos científicos, no se tendrá en cuenta escritos de introducciones, metodologías, resultados, discusiones y conclusiones.
- El sistema tendrá como corpus inicial no más de 50 resúmenes de textos científicos.
- El software será desarrollado en un plazo no mayor a 4 meses.

### 5.2 Riesgos

En esta sección, se presentará de forma detallada los posibles riesgos que se pueden presentar a lo largo de la ejecución del proyecto, así también, se mencionará su impacto sobre el proyecto, y las medidas correctivas para su mitigación.

Riesgo identificado	Impacto en el proyecto	Medidas correctivas para mitigar
No encontrar suficientes textos para formar el corpus de resúmenes.	Tener un corpus pequeño de entrenamiento tiene como consecuencia un aprendizaje deficiente y propenso a errores por la falta de generalización.	Conversar con personas con experiencia en el área para que facilite un poco la búsqueda de los textos para formar el corpus.
Elección de un algoritmo inadecuado para la clasificación de las secciones de un resumen.	Hay altas probabilidades de que el modelo no generalice bien y, por tanto, falle en las predicciones en los datos de prueba.	Realizar una investigación minuciosa sobre los algoritmos de clasificación utilizados en proyectos similares para lograr el mismo objetivo.
Falta de experiencia en las herramientas utilizadas.	Retraso en el proyecto debido a la curva de aprendizaje. Aprender a usar la herramienta puede tomar varios días, generando retrasos sustanciales.	Contar con tutoriales y manuales que muestren una explicación adecuada del uso de las herramientas para que puedan ser usados frente a cualquier duda.
Retraso en la presentación de los entregables acordados en los Sprints.	Genera retraso en el proyecto, ya que no se pueden realizar las tareas siguientes que dependen del entregable faltante.	Determinar una solución al problema y reestructurar las tareas. Sancionar en caso de incumplimiento por irresponsabilidad.

**Tabla 10: Riesgos del proyecto de fin de carrera.**

## 6 Justificación y viabilidad

### 6.1 Justificativa del proyecto de tesis

- Es necesario efectuar el estudio de las buenas prácticas realizadas para redactar resúmenes de textos científicos correctamente, ya que es una técnica básica y fundamental para poder redactar correctamente textos científicos más complejos. Debido a esto, se desea crear una herramienta que ayude a los estudiantes a redactar dichos resúmenes.
- La herramienta que se desarrollará a lo largo del proyecto, beneficiará a los universitarios e incluso estudiantes de nivel escolar u otro que estén interesados en el tema de la redacción de resúmenes de textos científicos. Gracias a esta herramienta, ellos podrán identificar sus errores y deficiencias en la redacción, y serán capaces de mejorar de forma autodidacta.
- Esta investigación, probablemente, puede servir de punto de partida para otros proyectos del mismo rubro con un alcance más profundo. Por ejemplo, otros proyectos podrían abarcar, no sólo resúmenes, sino también introducciones, metodologías, discusiones, conclusiones, entre otros; incluso podría ayudar a proyectos donde se desea implementar la herramienta en otro idioma.

### 6.2 Análisis de viabilidad del proyecto de tesis

Desde el punto de vista económico, el proyecto es considerado viable, pues éste es un proyecto sin fines de lucro. La herramienta que se desarrollará en a lo largo del proyecto, no está destinada a la venta hacia un público específico. Por otro lado, sólo

se hará uso de herramientas gratuitas, lo cual deja sin efecto cualquier limitación financiera. Los únicos gastos que se harán son los relacionados a la elaboración de la documentación, que son básicamente materiales de oficina.

El proyecto tendrá una duración de 4 meses. Este lapso de tiempo será dividido de manera eficaz para que se pueda culminar el proyecto en su totalidad, teniendo en cuenta la curva de aprendizaje de las herramientas, así como las actividades propias del proyecto. Estas actividades que se realizarán a lo largo del proyecto, y la duración de éstas, se pueden apreciar en la plan de actividades, mostrado en la **Tabla 11**. Teniendo como referencia lo descrito anteriormente, se puede decir que el proyecto también es viable desde el punto de vista temporal.

## 7 Plan de actividades

En esta sección, se mostrará el plan de actividades que se desarrollarán a lo largo del proyecto de fin de carrera, el cual contiene las actividades necesarias para implementar la totalidad de la solución y los tiempos estimados para llevar a cabo cada una de estas actividades.

Id	Nombre de tarea	Duración	Comienzo	Fin
1	<b>Proyecto de fin de carrera</b>	<b>104 días</b>	<b>mar 20/08/13</b>	<b>dom 01/12/13</b>
2	<b>Formar un corpus de resúmenes de texto científicos</b>	<b>23 días</b>	<b>mar 20/08/13</b>	<b>mié 11/09/13</b>
3	Revisar y actualizar los documentos previos realizados en Tesis 1	1 día	mar 20/08/13	mar 20/08/13
4	<b>Reunirse con el asesor - Presentar el entregable 1</b>	1 día	mié 21/08/13	mié 21/08/13
5	Realizar un manual de anotación (1)	4 días	jue 22/08/13	dom 25/08/13
6	<b>Exponer el entregable 1</b>	1 día	lun 26/08/13	lun 26/08/13
7	Realizar un manual de anotación (2)	1 día	mar 27/08/13	mar 27/08/13
8	<b>Reunirse con el asesor - Presentar el entregable 2</b>	1 día	mié 28/08/13	mié 28/08/13
9	Recopilar resúmenes de textos científicos, identificar sus componentes y realizar un análisis estadístico (1)	4 días	jue 29/08/13	dom 01/09/13
10	<b>Exponer el entregable 2</b>	1 día	lun 02/09/13	lun 02/09/13
11	Recopilar resúmenes de textos científicos, identificar sus componentes y realizar un análisis estadístico (2)	1 día	mar 03/09/13	mar 03/09/13
12	<b>Reunirse con el asesor - Presentar el entregable 3</b>	1 día	mié 04/09/13	mié 04/09/13
13	Recopilar resúmenes de textos científicos, identificar sus componentes y realizar un análisis estadístico (3)	4 días	jue 05/09/13	dom 08/09/13
14	<b>Exponer el entregable 3</b>	1 día	lun 09/09/13	lun 09/09/13
15	Recopilar resúmenes de textos científicos, identificar sus componentes y realizar un análisis estadístico (4)	1 día	mar 10/09/13	mar 10/09/13
16	<b>Reunirse con el asesor - Presentar el entregable 4</b>	1 día	mié 11/09/13	mié 11/09/13
17	<b>Determinar las características que serán extraídas de cada oración de los resúmenes de textos científicos</b>	<b>28 días</b>	<b>jue 12/09/13</b>	<b>mié 09/10/13</b>
18	Recopilar información sobre las características más importantes del resumen de texto científico (1)	4 días	jue 12/09/13	dom 15/09/13
19	<b>Exponer el entregable 4</b>	1 día	lun 16/09/13	lun 16/09/13
20	Recopilar información sobre las características más importantes del resumen de texto científico (2)	1 día	mar 17/09/13	mar 17/09/13
21	<b>Reunirse con el asesor - Presentar el entregable 5</b>	1 día	mié 18/09/13	mié 18/09/13
22	Determinar las características que serán extraídas de cada resumen de texto científico (1)	4 días	jue 19/09/13	dom 22/09/13
23	<b>Exponer el entregable 5</b>	1 día	lun 23/09/13	lun 23/09/13
24	Determinar las características que serán extraídas de cada resumen de texto científico (2)	1 día	mar 24/09/13	mar 24/09/13
25	<b>Reunirse con el asesor - Presentar el entregable 6</b>	1 día	mié 25/09/13	mié 25/09/13
26	Implementar un algoritmo que se encargue de la extracción de las características de los resúmenes (1)	4 días	jue 26/09/13	dom 29/09/13
27	<b>Exponer el entregable 6</b>	1 día	lun 30/09/13	lun 30/09/13



28	Implementar un algoritmo que se encargue de la extracción de las características de los resúmenes (2)	1 día	mar 01/10/13	mar 01/10/13
29	<b>Reunirse con el asesor</b>	1 día	mié 02/10/13	mié 02/10/13
30	Implementar un algoritmo que se encargue de la extracción de las características de los resúmenes (3)	6 días	jue 03/10/13	mar 08/10/13
31	<b>Reunirse con el asesor</b>	1 día	mié 09/10/13	mié 09/10/13
32	<b>Implementar un modelo clasificador y evaluar su desempeño utilizando métricas tales como Precision, Recall y F-Measure.</b>	<b>19 días</b>	<b>jue 10/10/13</b>	<b>lun 28/10/13</b>
33	Implementar un modelo clasificador y evaluar su desempeño utilizando métricas tales como Precision, Recall y F-Measure; si éste es malo, se procede a seleccionar mejores características (1)	6 días	jue 10/10/13	mar 15/10/13
34	<b>Reunirse con el asesor - Presentar el entregable parcial</b>	1 día	mié 16/10/13	mié 16/10/13
35	Implementar un modelo clasificador y evaluar su desempeño utilizando métricas tales como Precision, Recall y F-Measure; si éste es malo, se procede a seleccionar mejores características (2)	6 días	jue 17/10/13	mar 22/10/13
36	<b>Reunirse con el asesor</b>	1 día	mié 23/10/13	mié 23/10/13
37	Implementar un modelo clasificador y evaluar su desempeño utilizando métricas tales como Precision, Recall y F-Measure; si éste es malo, se procede a seleccionar mejores características (3)	4 días	jue 24/10/13	dom 27/10/13
38	<b>Exposición Parcial</b>	1 día	lun 28/10/13	lun 28/10/13
39	<b>Implementar una aplicación que clasifique automáticamente las oraciones de los resúmenes de textos científicos, siguiendo una estructura pre-definida</b>	<b>33 días</b>	<b>mar 29/10/13</b>	<b>sáb 30/11/13</b>
40	Implementar un modelo clasificador y evaluar su desempeño utilizando métricas tales como Precision, Recall y F-Measure; si éste es malo, se procede a seleccionar mejores características (4)	1 día	mar 29/10/13	mar 29/10/13
41	<b>Reunirse con el asesor</b>	1 día	mié 30/10/13	mié 30/10/13
42	Implementar un modelo clasificador y evaluar su desempeño utilizando métricas tales como Precision, Recall y F-Measure; si éste es malo, se procede a seleccionar mejores características (5)	3 días	jue 31/10/13	sáb 02/11/13
43	Desarrollar la interfaz gráfica y aplicación de clasificación de componentes (1)	3 días	dom 03/11/13	mar 05/11/13
44	<b>Reunirse con el asesor</b>	1 día	mié 06/11/13	mié 06/11/13
45	Desarrollar la interfaz gráfica y aplicación de clasificación de componentes (2)	4 días	jue 07/11/13	dom 10/11/13
46	<b>Presentar el entregable final</b>	1 día	lun 11/11/13	lun 11/11/13
47	Desarrollar la interfaz gráfica y aplicación de clasificación de componentes (3)	1 día	mar 12/11/13	mar 12/11/13
48	<b>Reunirse con el asesor</b>	1 día	mié 13/11/13	mié 13/11/13
49	Desarrollar la interfaz gráfica y aplicación de clasificación de componentes (4)	6 días	jue 14/11/13	mar 19/11/13
50	<b>Reunirse con el asesor</b>	1 día	mié 20/11/13	mié 20/11/13
51	Revisar que el software esté funcionando correctamente y corregir los errores detectados	9 días	jue 21/11/13	vie 29/11/13
52	<b>Exposición Final</b>	1 día	sáb 30/11/13	sáb 30/11/13
53	<b>Cierre del proyecto</b>	1 día	dom 01/12/13	dom 01/12/13

**Tabla 11: Plan de actividades que se realizarán a lo largo del proyecto de fin de carrera.**

## CAPÍTULO 3: Corpus de Resúmenes

### 1 Breve descripción

En esta sección, se presentará una serie de estadísticas, las cuales fueron obtenidas del análisis del corpus de textos recopilado. Estas ayudarán a apreciar claramente la situación actual de los alumnos de la PUCP en comparación con las demás universidades, en cuanto a la redacción de este tipo de textos, lo que nos permite tomar decisiones para mejorar esta situación. Así también, se podrá ver la frecuencia de las diferentes categorías retóricas (contexto, brecha, propósito, metodología, resultado, conclusión y estructura) en el corpus de resúmenes científicos.

### 2 Corpus vs Áreas de Informática

Los textos del corpus fueron recolectados desde el repositorio de tesis de la PUCP. Se enfocó en recolectar tesis de pregrado, publicadas desde el 2010 hasta el 2013. Éstas fueron escritas por estudiantes de la facultad de Ingeniería de Informática de la PUCP. Así también, se añadieron tesis doctorales, publicadas desde el 2006 hasta el 2013, de universidades extranjeras tales como la Universidad Autónoma de Barcelona, la Universidad Politécnica de Catalunya, entre otras. Los textos del corpus abarcan diferentes áreas de Informática; por tal motivo, se decidió dividirlos en 5 áreas: Sistemas de información, Tecnologías de información, Ingeniería de Software, Ciencias de la Computación e Ingeniería de Computadoras. En la **Tabla 12**, se puede apreciar la cantidad de resúmenes clasificados en cada área, así como el número total y promedio de palabras de los resúmenes.

Área de Informática	Tesis de pregrado de la PUCP	Tesis de posgrado de diferentes universidades extranjeras
Sistemas de Información	14	-
Tecnologías de Información	6	2
Ingeniería de Software	6	-
Ciencias de la Computación	4	12
Ingeniería de Computadoras	-	-
<b>Total</b>	<b>44</b>	<b>14</b>

**Tabla 12: Distribución del corpus a través de las áreas de Informática.**

Para realizar este análisis, se emplearon 30 tesis de pregrado de la PUCP y 14 tesis de posgrado de universidades extranjeras. Como se había mencionado anteriormente, en la **Tabla 12**, se puede apreciar la división de estos textos, tanto de pregrado como posgrado, en las diversas áreas de Informática.

En la PUCP, se da mayor énfasis al área de Sistemas de Información, sin embargo, en la **Tabla 12**, se puede apreciar que la mayor cantidad de tesis pertenecen al área de Ciencias de Computación. Esto se debe a que cerca del 90% de las tesis de universidades extranjeras pertenecen a esta área.

### 3 Número de componentes por resumen

Considerando los componentes del esquema de anotación presentado en el Manual de Anotación (ver Anexo 1), solo 3 (6.8%) de los resúmenes revisados contiene los seis primeros componentes del esquema. Con respecto a los demás casos, de acuerdo a lo señalado en la **Tabla 13**, se puede ver que 8 (18.2%) de los resúmenes presentan 5 componentes, 11 (25%) presentan 4 componentes, 11 (25%) presentan 3 componentes y 11 (25%) presentan 2 componentes.

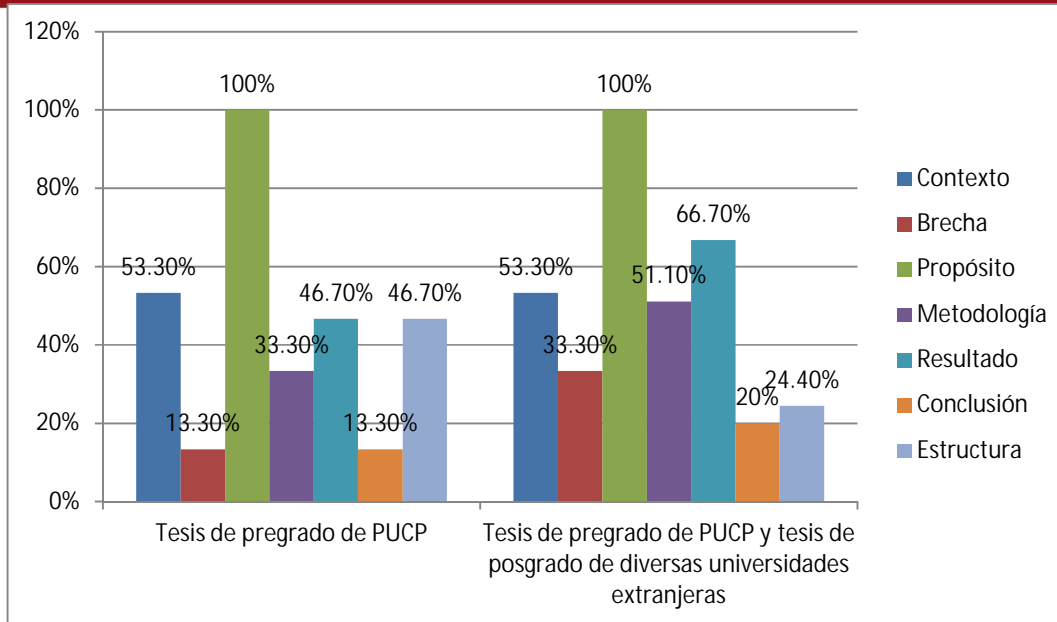
Área de Informática	Número de componentes					
	1	2	3	4	5	6
Sistemas de Información	-	5	2	4	2	-
Tecnologías de Información	-	3	2	2	-	1
Ingeniería de Software	-	1	3	1	1	-
Ciencias de la Computación	-	2	4	4	5	2
Ingeniería de Computadoras	-	-	-	-	-	-
<b>Total</b>	-	<b>11</b>	<b>11</b>	<b>11</b>	<b>8</b>	<b>3</b>

**Tabla 13: Número de componentes por resumen.**

Se puede observar que 3 resúmenes contienen las 6 categorías, estos textos pertenecen a los estudiantes de posgrado. Así también, se puede ver que en el área de las Ciencias de la Computación, una buena parte de los resúmenes contienen desde 4 a 6 categorías. Esto se debe a que cerca del 90% de las tesis de posgrado pertenecen a esta área. Teniendo como base este análisis, se puede decir que los alumnos de posgrado tienen una idea clara de cómo desarrollar resúmenes de textos científicos, mientras que los de pregrado no. Este es el resultado que se esperaba, pues los estudiantes de posgrado se encuentran mejor preparados, ya que es muy probable que muchos de ellos ya tengan una gran experiencia en la redacción de este tipo de textos, debido a la necesidad de escribir artículos científicos, informes, entre otros.

### 4 Frecuencia de los componentes en el corpus

La frecuencia de cada categoría del esquema de anotación en el corpus está presentada en la **Figura 9**. Se puede apreciar 2 gráficos, uno de ellos está asociado a las tesis de pregrado de la facultad de Informática de la PUCP, y el otro a las tesis de posgrado de la facultad de Informática de diversas universidades extranjeras, mencionadas en el acápite 1 de este capítulo.



**Figura 9: Distribución de las categorías en el corpus**

En la **Figura 10**, en el gráfico asociado a la tesis de pregrado de la PUCP, se puede observar que el porcentaje de aparición de la categoría “Estructura” es muy alto en comparación del gráfico asociado al total de tesis del corpus. Teniendo en cuenta que la categoría “Estructura” no debería pertenecer a un resumen de texto científico, se puede inferir que los alumnos de pregrado de la facultad de Informática de la PUCP no tienen claro la estructura de un resumen académico, pues consideran que dicha categoría forma parte del resumen, lo cual no es cierto.

Por otro lado, se puede ver que al agregar los resúmenes de los estudiantes de posgrado, el porcentaje aparición de las categorías en los textos mejoró en general. En el caso de “Propósito”, tanto en la tesis de pregrado como de posgrado, se mantuvieron en un 100%, lo cual indica que todas las tesis mencionan cual es el objetivo principal de sus proyectos. Por otro lado, el porcentaje de aparición de la categoría “Estructura” en los textos disminuyó considerablemente, lo cual indica que los estudiantes de posgrado tienen claro cuál es la estructura de un resumen académico, ya que no usan esta categoría en sus textos. En el caso de las demás categorías, sus porcentajes aumentaron bastante, lo cual es una buena señal, pues se puede notar que en los textos de los estudiantes de posgrado están presentes la gran mayoría de las categorías.

## CAPÍTULO 4: Conjunto de atributos que se emplearán para la identificación de las categorías

### 1 Breve descripción

En la **Tabla 14**, se puede observar el esquema de clasificación de oraciones de los resúmenes de textos científicos. Cada una de las categorías está asociada a una etiqueta (primeras letras del nombre de la categoría **en inglés**). Los atributos que se extraerán tienen como objetivo la clasificación de las oraciones del resumen en estas categorías.

Etiqueta	Categoría (Inglés)	Categoría (Español)	Descripción
B	Background	Contexto	Conocimiento aceptado por la comunidad científica
G	Gap	Brecha	Problema de investigación, necesidades, ...
P	Purpose	Propósito	Propósito de la investigación
M	Methodology	Metodología	Metodología utilizada
R	Result	Resultado	Resultados obtenidos
C	Conclusion	Conclusión	Conclusión, recomendación, contribución, ...
S	Structure	Estructura	Descripción de las partes del texto / asuntos tratados

**Tabla 14: Esquema de clasificación.**

### 2 Visión general de los atributos

El AZEsp es un clasificador que atribuirá a cada oración de entrada una posible categoría retórica de la lista mencionada en la **Tabla 14**. Así como otros algoritmos de aprendizaje de máquina, en vez de lidiar directamente con el objeto que será clasificado, el AZEsp recibirá las oraciones como vectores de atributos. Debido a esto, la extracción de atributos es un paso crucial en tal escenario.

En la **Tabla 15**, se muestra una descripción resumida del conjunto de atributos que utilizará AZEsp con el objetivo de clasificar las oraciones de los resúmenes.

Atributo	Descripción	Valores posibles
1. Tamaño	¿Cuál es el tamaño de la oración? (en base a los límites de 20 y 40 palabras)	Corta, media o larga
2. Localización	¿Cuál es la posición de la oración en el resumen?	Primera, segunda, mediana, penúltima o última
3. Expresión	¿A qué categoría retórica pertenece la expresión estándar contenida en la oración?	B, G, P, M, R, C, S o <i>noexpr</i>
4. Tiempo	¿Cuál es el tiempo del primer verbo finito de la oración?	IMP, PRES, PAST, FUT, COND, PRES-CPO, PAST-CPO, FUT-CPO, PRES-CT, PAST-CT, FUT-CT, PRES-CPO-CT, PAST-CPO-CT, FUT-CPO-CT o <i>noverb</i>
5. Voz	¿Cuál es la voz del primer verbo finito de la oración?	Pasiva, activa o <i>noverb</i>

6. Modal	¿El primer verbo finito de la oración es modal?	Sí, no o <i>noverb</i>
7. Histórico	¿Cuál es la categoría de la oración anterior?	_, B, G, P, M, R, C o S

**Tabla 15: Resumen del conjunto de atributos.**

### 3 Descripción detallada de los atributos

En esta sección, se describirá, en mayor detalle, los atributos anteriormente mencionados en la **Tabla 15**.

#### 3.1 Tamaño

El atributo **Tamaño** clasifica una oración como *corta*, *mediana* o *larga*, basado en el número de palabras de la oración. Para determinar el valor del atributo son utilizados los límites de 20 y 40 palabras. Teniendo esto como base, se cumple lo siguiente:

- La oración es *corta* si el número de palabras es menor a 20.
- La oración es *mediana* si el número de palabras es mayor a 20 y menor a 40.
- La oración es *larga* si el número de palabras es mayor a 40.

Cabe resaltar que esos límites fueron estimados teniendo como base la media de los tamaños de las oraciones presentes en el corpus.

#### 3.2 Localización

El atributo **Localización** identifica la posición ocupada por la oración en el resumen. Las separaciones hechas por los párrafos no fueron consideradas, todo resumen fue procesado como si fuera un párrafo único. Para este atributo, son utilizados cinco valores: *primera*, *segunda*, *mediana*, *penúltima* y *última*. Para la determinación de los valores de **Localización** se utilizaron los cinco valores descritos anteriormente, los cuales caracterizan localizaciones comunes para algunas de las categorías del esquema utilizado.

#### 3.3 Expresión

El atributo **Expresión** identifica la presencia de una expresión “estándar” en la oración. El método utilizado para encontrar dichas expresiones es el reconocimiento de estándares en base a un conjunto fijo de expresiones que pueden aparecer en textos que siguen el modelo de estructuración presentado en la **Tabla 14**. Teniendo esto en cuenta, nuestro conjunto de expresiones es dividido de acuerdo a las categorías previstas en este modelo. En la **Tabla 16**, se presenta ejemplos de los tipos de expresiones representadas por el atributo **Expresión**.

Categoría	Expresión
Contexto (B)	A partir del año...
Brecha (G)	Sin embargo, es necesario...
Propósito (P)	Esta tesis presenta...
Metodología (M)	Fue utilizado el modelo...
Resultado (R)	Los resultados muestran...
Conclusión (C)	Se concluye que...
Estructura (S)	En la sección siguiente...

**Tabla 16: Ejemplos de expresiones estándares.**

### 3.4 Tiempo, Voz y Modal

Los atributos **Tiempo**, **Voz** y **Modal**, llamados también atributos sintácticos, describen propiedades sintácticas del primer verbo finito de la oración, en modo indicativo o imperativo. Debido a la alta probabilidad de verbos en modo subjuntivo pertenecientes a oraciones subordinadas, ellos son considerados solo cuando ningún otro verbo finito en indicativo o imperativo es encontrado. En el caso de que ningún verbo finito sea encontrado en la oración, los tres atributos sintácticos toman el valor de *noverb*. Es importante destacar que la determinación de los atributos sintácticos son considerados tanto verbos simples (Ejm: "Los resultados demuestran..."), como locuciones verbales, llamadas también como verbos complejos, que incluyen uno o más verbos auxiliares para expresar lo siguiente:

- i. Aspecto continuo (*estar* + gerundio), o perfecto (*hacer* + participio), como en "Este gran trabajo ha sido realizado para..."
- ii. Voz pasiva (*ser* + participio), como en "ha sido realizado"
- iii. Modalización (*deber/poder/precisar/tener (que)etc.* + infinitivo).

El verbo complejo también puede contener el pronombre *se* como índice de indeterminación del sujeto o partícula pasiva. En este trabajo, será utilizado el término "frase verbal" para designar verbos en general, sean los simples o complejos.

El atributo **Tiempo** indica la flexión del verbo (simple o complejo) y puede asumir 14 valores, incluido el valor *noverb*. NOVERB para frases sin verbo, IMP para frases imperativas, o algún identificador en el formato "*SimpleTense-(not)perfect-(not)continuous*", donde "*SimpleTense*" indica el tiempo del componente finito en la frase verbal, "*(not)perfect*" indica la presencia del verbo auxiliar *haber* en la frase verbal expresando el aspecto perfecto, y "*(not)continuous*" indica la presencia de verbo auxiliar *estar* expresando el aspecto continuo. Puede ver ejemplos de estos identificadores en la **Tabla 15**.

El atributo **Voz** indica la voz del verbo, y puede asumir los valores *activo*, *pasivo* o *noverb*. La voz pasiva es entendida en un sentido más amplio, albergando ciertas formas y construcciones verbales que son generalmente utilizadas para omitir un agente, es decir:

- i. Voz pasiva analítica (verbo *ser* + participio)
- ii. Voz pasiva sintética (hecha con la partícula pasiva *se*)
- iii. Sujeto indeterminado indicado por la flexión de tercera persona en singular (hecha con la partícula pasiva *se*)

El atributo **Modal** indica si hay un auxiliar modal en la frase verbal y puede asumir los valores *sí*, *no* y *noverb*. Son considerados como modales los siguientes verbos: *tener* (*que*), *deber* y *poder*.

### 3.5 Histórico

El atributo **Histórico** toma en cuenta la categoría de la oración anterior a la oración que está siendo clasificada. De acuerdo a lo dicho anteriormente, se sabe que algunas zonas argumentativas tienden a seguir otras zonas específicas. En el corpus, algunas secuencias de categorías son muy frecuentes. Por ejemplo, el patrón de Contexto (B) seguido de Brecha (G), con repetición o no, y este seguido del Propósito (P), es decir, ((BG)|(GB)+)P, ocurre en una gran parte del corpus.

Durante el entrenamiento, la determinación del atributo **Histórico** es hecha por una serie de observaciones del corpus. Para las oraciones nuevas, no obstante, el atributo **Histórico** tiene que ser estimado como un segundo paso durante el proceso de testeo.





## CAPÍTULO 5: Extracción de los valores de los atributos

### 1 Breve descripción

Los atributos descritos anteriormente son determinados automáticamente a partir del texto de entrada, por medio de un proceso implementado en Java. En la **Figura 10**, se puede apreciar las etapas del proceso de determinación de los atributos, así como los recursos necesarios para poder llevar a cabo cada etapa. Así también, se muestra los atributos resultantes de cada etapa.

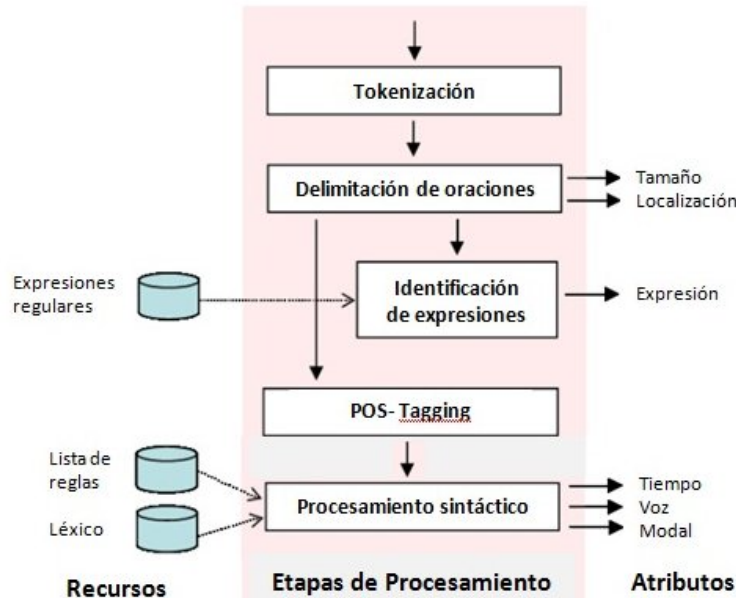


Figura 10: Etapas del proceso de determinación de atributos.

### 2 Proceso de determinación de atributos

A continuación, se detallará cada una de las etapas del proceso de determinación de atributos, que se emplearán para la identificación de las categorías de cada oración.

#### 2.1 Tokenización y Delimitación de oraciones

Para la realización de las primeras etapas del proceso, **Tokenización y Delimitación de oración**, se utilizó la herramienta *Freeling*, descrita anteriormente en la sección de “Herramientas, métodos y procedimientos” (ver **Capítulo 2**).

En la etapa de **Tokenización**, se divide el texto en unidades independientes más pequeñas, es decir, las palabras. De la librería *Freeling*, se empleó la clase *tokenizer*, el cual recibe un texto plano y retorna una lista de objetos *word*.

Posteriormente, en la etapa de **Delimitación de oraciones**, se agrupan las palabras obtenidas en la primera etapa con el fin de obtener las oraciones del texto. De la librería *Freeling*, se empleó la clase *splitter*, el cual recibe la lista de objetos *word*,

obtenida previamente, y retorna una lista de objetos *sentence*. Esta segunda etapa provee la información necesaria para obtener los atributos **Tamaño** y **Localización**.

Por otro lado, cabe resaltar que no hubo problemas con la herramienta, en cuanto a la presencia de paréntesis, corchetes y llaves, así como la presencia de puntos debido a las abreviaturas (Ejemplo: Dr., Sr.), los cuales pueden ser causantes de malas interpretaciones, pues puede ser considerada como punto final de una oración.

## 2.2 Identificación de expresiones

Para hacer el reconocimiento de las expresiones estándares, se construyó un conjunto de expresiones estándares, divididas en seis categorías: Contexto, Brecha, Propósito, Metodología, Resultado y Conclusión. Luego, se construyó un algoritmo, donde se empleó dicha lista de expresiones, para la identificación de éstas en cada oración del texto. Si la expresión contenida en la oración pertenecía a la categoría C, entonces el valor del atributo **Expresión** sería C. En la **Tabla 17**, se puede ver una parte de la lista de las expresiones estándares identificadas en el corpus de resúmenes científicos (Para mayor detalle, ver Anexo B).

Categoría	Expresión
Contexto (B)	En el transcurrir de las últimas décadas... Para el primer trimestre... En la actualidad... Hoy en día... En el ambiente de negocios de hoy... En los últimos años...
Brecha (G)	Sin embargo... ...la problemática actual... No obstante...
Propósito (P)	El tema de tesis tiene como objetivo... El presente trabajo de tesis implementa... El presente trabajo de tesis presenta...
Metodología (M)	Se emplearán metodologías... El análisis... El diseño... La implementación...
Resultado (R)	La solución consiste en... A partir de los resultados... ...tiene las siguientes características...
Conclusión (C)	Se tiene como trabajo futuro... Este trabajo contribuye... Se registran como conclusiones...
Estructura (S)	La estructura de la presente tesis contiene... El presente documento ha sido estructurado... En el primer capítulo...

**Tabla 17: Ejemplos de expresiones estándares encontradas en el corpus de resúmenes científicos.**

## 2.3 POS-Tagging

La etiquetación morfosintáctica, o *POS-Tagging*, es una etapa fundamental en el proceso de determinación de atributos, pues brinda información relevante para el procesamiento sintáctico de las palabras del texto.

De la librería *Freeling*, se empleó la clase *maco*, el cual recibe una lista que recibe una lista de objetos *sentence*, y anota morfológicamente cada objeto *word* de cada oración dada. Incluye sub-módulos, tales como detección de días, números, etc.

## 2.4 Procesamiento sintáctico

Gran parte del esfuerzo de implementación de la extracción de atributos se aplicó a los atributos relacionados a los verbos: **Tiempo**, **Voz** y **Modal**. Esto es debido, en parte, a la gran flexibilidad morfológica del español y, por otro lado, debido a que la herramienta *Freeling* tiene ciertas limitaciones.

La librería *Freeling* ayuda a clasificar las palabras en: adjetivos, adverbios, determinantes, nombres, verbos, pronombres, conjunciones, interjecciones, preposiciones, entre otros. Para la obtención de los atributos mencionados anteriormente, se debe enfocar en las palabras clasificadas como verbos. Aquí se generan algunos inconvenientes, ya que la herramienta no brinda funcionalidades que ayuden a la identificación de algunos tiempos verbales, algunos modales y la voz pasiva.

En cuanto a los tiempos verbales, se tuvo que crear un algoritmo para la identificación de los verbos en tiempo pretérito perfecto compuesto, futuro compuesto, entre otros, en su mayoría verbos compuestos por dos o más palabras. En cuanto a los modales, la herramienta no los identifica, sin embargo, se hizo un pequeño algoritmo, donde se empleó una lista de modales previamente identificados en el corpus, para la identificación de éstos. Finalmente, en cuanto a la voz pasiva, se tuvo el mismo problema que los tiempos verbales: la identificación de verbos compuestos; pues en cuanto a la voz pasiva analítica era necesario identificar el verbo ser acompañado de un verbo participio. Así también, hubo ciertos problemas con la identificación el pronombre *se*. En todos los casos, como se mencionó anteriormente, se tuvo que hacer ciertos algoritmos para poder obtener los valores de los atributos **Tiempo**, **Voz** y **Modal**, tal como se describió en el capítulo anterior.

De la librería *Freeling*, se empleó la clase *tagger*, el cual recibe una lista de objetos *sentence*, y etiqueta gramaticalmente y morfosintácticamente cada objeto *word* de cada oración dada.

En la **Tabla 18**, se puede apreciar la etiquetación que da *Freeling* a los verbos. Para la determinación del tiempo verbal, modal y voz del verbo, se tuvo como base lo señalado en dicha tabla. Luego, se hizo una serie de algoritmos, con el fin de obtener los valores de los atributos **Tiempo**, **Voz** y **Modal**, solucionando los problemas descritos anteriormente.

VERBOS			
Pos.	Atributo	Valor	Código
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A
		Semiauxiliar	S
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Infinitivo	N
		Gerundio	G
		Participio	P
4	Tiempo	Presente	P
		Imperfecto	I
		Futuro	F
		Pasado	S
		Condicional	C
		-	0
5	Persona	Primera	1
		Segunda	2
		Tercera	3
6	Número	Singular	S
		Plural	P
7	Género	Masculino	M
		Femenino	F

Tabla 18: Categorías de la etiquetación de los verbos. Fuente: Freeling, *s/f*.

En la **Tabla 19**, se puede ver un ejemplo de etiquetación de las palabras identificadas como verbos (simples).

Forma	Lema	Etiqueta
cantada	cantar	VMP00S <i>F</i>
cantadas	cantar	VMP00P <i>F</i>
cantado	cantar	VMP00S <i>M</i>
cantados	cantar	VMP00P <i>M</i>

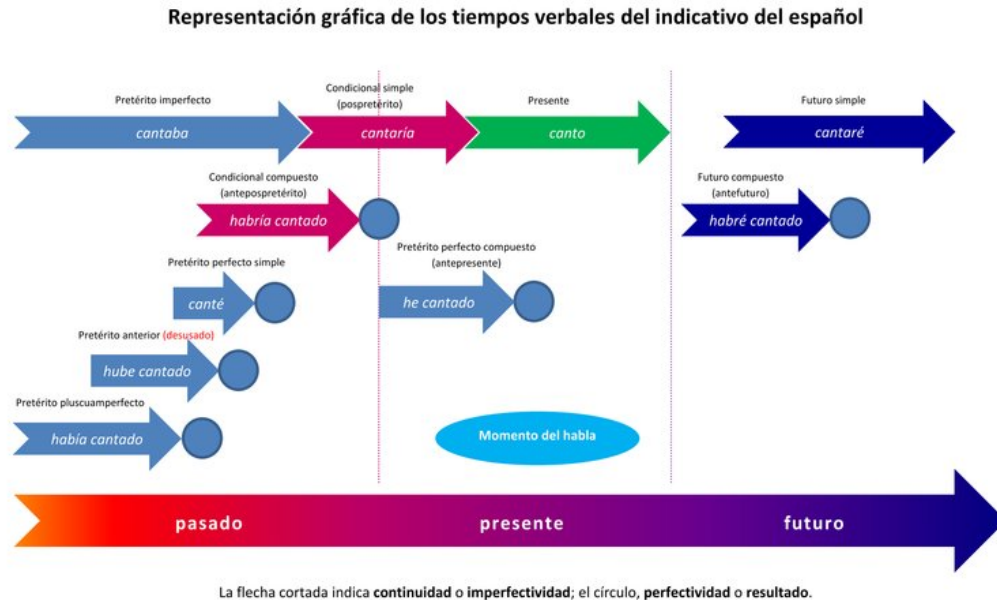
Tabla 19: Ejemplo de etiquetación de los verbos. Fuente: Freeling, *s/f*.

En cuanto a la etiquetación, hecha por *Freeling*, se debe tener en cuenta los siguientes aspectos:

- El lema del verbo siempre será infinitivo.
- Se etiqueta las formas del verbo *haber* como auxiliares (VA) cuando actúan como tal, y como verbo principal (VM) en los existenciales (Ejemplo: hay dinero, cuando haya dinero).
- Se etiqueta las formas del verbo *ser* como semiauxiliares (VS).
- Se etiqueta los verbos restantes como principales (VM).
- El atributo *Género* solo afecta a los participios, para el resto de formas este atributo no se especifica (0).

- Para las formas de infinitivo y gerundio no se especifican los atributos de Tiempo, Persona, Número y Género, por lo que su valor será 0.

Para el caso de los verbos compuestos, se tuvo que usar un algoritmo específico que los pueda identificar. Para poder explicar la lógica de éste, se tendrá como base la **Figura 11**.



**Figura 11: Ejemplo de verbos simples y compuestos. Fuente: Wikipedia, 2013.**

Por ejemplo, en el caso del pretérito perfecto compuesto (*he cantado*), suponiendo que dicho verbo fue el primero de la oración analizada, se identificó la presencia de los verbos *haber* y *cantar*. Luego, con la ayuda de *Freeling*, se obtuvo el código de ambos; en el caso de *haber* se tiene **VAIP1SM** y el de *cantar*, **VMP00SM**. Se presta atención a cierta parte de éste para determinar el valor del atributo **Tiempo**. El hecho de ver un **VA** seguido de un **VM**, señala que es un verbo compuesto; sin embargo, el análisis no acaba ahí. Para determinar el tiempo del verbo compuesto, se debe prestar atención al tiempo del primer verbo (*haber*), el cual es Presente (**VAIP1SM**); así también se debe observar el Modo del segundo verbo (*cantar*), el cual es Participio (**VMP00SM**). Entonces, una vez obtenido ambos valores, se puede decir que el verbo compuesto tiene **PRES-CPO** como valor del atributo **Tiempo**. Para mayor información sobre los valores del atributo **Tiempo**, ver la **Tabla 15**, así como su descripción detallada en el Capítulo 4.

Para calcular los demás valores, se hace un análisis similar al mostrado anteriormente. De la misma forma, se determina el valor del atributo **Voz** y **Modal**. En el caso del **Modal**, se utilizará adicionalmente una lista de modales pre-definidos (Ver Capítulo 4).

## CAPÍTULO 6: Evaluación del Modelo Clasificador

### 1 Breve descripción del clasificador estadístico

Se utilizó el modelo clasificador de Naive Bayes para estimar las probabilidades que una oración *S* tenga una categoría *C*, teniendo como base el valor de sus atributos. La categoría *C* que tenga una mayor probabilidad es escogida como salida para la oración *S*.

Como se sabe, el aprendizaje es supervisado; es decir, durante la fase de entrenamiento, el clasificador aprende las asociaciones entre los atributos y las categorías proveniente del corpus manualmente anotado. Durante la fase de testeo, el modelo previamente entrenado proporciona una probabilidad de cada categoría para cada oración de entrada, en base a los atributos identificados en dicha oración.

### 2 Evaluación del clasificador

Se empleó un corpus de 44 resúmenes para realizar esta evaluación. Se hicieron 2 experimentos para medir la capacidad de clasificación del modelo. En el primero, se consideran todas las categorías definidas (Contexto, Brecha, Propósito, Metodología, Resultado, Conclusión, Estructura). En el segundo caso, no se toma en cuenta la categoría "Estructura", ya que esta no debería formar parte de los resúmenes de textos científicos, según la literatura revisada.

Para la obtención de estos resultados se utilizó WEKA. Como se había mencionado anteriormente, se utilizó el clasificador de *Naive Bayes*. Así también, se aplicó *11-fold cross validation*; es decir, en cada iteración, el clasificador era entrenado con 40 resúmenes y testado con 4 resúmenes.

Más adelante, se mostrarán los resultados obtenidos en ambos experimentos. Dentro de estos resultados, en la sección *Summary*, se debe prestar atención al valor del *Kappa statistic*, ya que éste mide el nivel de concordancia que hay entre la clasificación hecha entre el anotador humano y clasificador AZEsp. Así también, es importante observar el valor de *Correctly Classified Instances*, pues éste representa la cantidad de oraciones *S* clasificadas correctamente, así como el porcentaje de oraciones clasificadas correctamente sobre el total de oraciones.

Por otro lado, para analizar el comportamiento de AZEsp en relación a las categorías, se debe prestar atención a las medidas *Precision*, *Recall* y *F-Measure*, ubicadas en la sección de *Detailed Accuracy By Class*. Para este caso, dada una categoría *C*, *Precision* es el total de oraciones correctamente clasificadas como *C* sobre el total de oraciones clasificadas como *C*. *Recall* es el total de oraciones correctamente clasificadas como *C* sobre el total de oraciones pertenecientes a la categoría *C* presentes en el conjunto.

En cuanto a la sección *Confusion Matrix*, los resultados dados por el clasificador están representados por las columnas, mientras que los resultados dados por el anotador humano están representados por las filas.

En la **Figura 12**, se puede observar los resultados obtenidos del Experimento 1, donde se toma en cuenta todas las categorías identificadas en el corpus.

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      184          40.2626 %
Incorrectly Classified Instances    273          59.7374 %
Kappa statistic                    0.2643
Mean absolute error                 0.2091
Root mean squared error             0.3238
Relative absolute error             88.1912 %
Root relative squared error         94.0705 %
Total Number of Instances          457

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
          0.403   0.051   0.574     0.403   0.474     0.761     B
          0         0.005   0         0         0         0.609     G
          0.333   0.04    0.556   0.333   0.417     0.712     P
          0.443   0.077   0.635   0.443   0.522     0.718     M
          0.284   0.165   0.271   0.284   0.277     0.639     R
          0.056   0.005   0.333   0.056   0.095     0.773     C
          0.725   0.393   0.314   0.725   0.439     0.7       S
Weighted Avg.  0.403   0.139   0.428   0.403   0.385     0.7

=== Confusion Matrix ===

 a  b  c  d  e  f  g  <-- classified as
27  1 12  3  6  0 18 | a = B
 4  0  1  3  7  0 19 | b = G
 6  1 20  3  8  0 22 | c = P
 1  0  1 47 16  0 41 | d = M
 6  0  1 12 23  1 38 | e = R
 0  0  0  4  7  1  6 | f = C
 3  0  1  2 18  1 66 | g = S
    
```

Figura 12: Resultados del Experimento 1, obtenidos por WEKA.

Se puede observar que el clasificador tiene un desempeño realmente malo para el caso de la categoría *Brecha* (G) ( $F\text{-Measure}=0$ ) y *Conclusion* (C) ( $F\text{-Measure}=0.095$ ). Se esperaba un resultado así, ya que, en el corpus, eran muy pocas las oraciones que fueron identificadas, manualmente, como *Brecha*. Muchos algoritmos de aprendizaje, incluido el Naive Bayes, tendrían un desempeño pésimo con categorías poco frecuentes, pues no se tiene suficiente material de entrenamiento para éstas. No obstante, con respecto a las demás categorías, tales como *Metodología* (M) ( $F\text{-Measure}=0.522$ ) y *Contexto* (B) ( $F\text{-Measure}=0.474$ ), se tiene un mejor desempeño del clasificador, este es debido a que se tienen más oraciones identificadas manualmente en esta categoría.

En general, se puede ver que el porcentaje de oraciones clasificadas correctamente por el clasificador es de un 40.26%, el cual es un poco bajo. Esto se debe a que aún no se ha implementado el extractor del atributo **Expresion**, el cual, según la literatura revisada, es el atributo con mayor poder de distinción. Así también, el atributo **Histórico**, el cual es el segundo más importante, tampoco ha sido implementado hasta el momento. Según la literatura revisada, este último es una característica bastante útil y mejoraría significativamente el desempeño del clasificador.

Sin embargo, se ha intentado mejorar esta situación en el Experimento 2, eliminando la categoría *Estructura* (S), la cual no debería pertenecer a un resumen, no obstante, es una de las categorías que más se presenta en el corpus. Los resultados se pueden apreciar en la **Figura 13**.

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      162          44.2623 %
Incorrectly Classified Instances    204          55.7377 %
Kappa statistic                    0.2636
Mean absolute error                0.1972
Root mean squared error            0.3152
Relative absolute error            86.6081 %
Root relative squared error        93.4776 %
Total Number of Instances          366

=== Detailed Accuracy By Class ===

          TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area   Class
          0.388     0.057     0.605     0.388     0.473     0.771     B
          0         0.012     0         0         0         0.636     G
          0.333     0.049     0.571     0.333     0.421     0.686     P
          0.745     0.4       0.432     0.745     0.547     0.749     M
          0.444     0.221     0.364     0.444     0.4       0.661     R
          0.056     0.003     0.5       0.056     0.1       0.757     C
          0         0         0         0         0         ?         S
Weighted Avg.   0.443     0.184     0.434     0.443     0.407     0.713

=== Confusion Matrix ===

 a  b  c  d  e  f  g  <-- classified as
26  2 12 20  7  0  0  | a = B
 4  0  1 21  8  0  0  | b = G
 6  1 20 22 11  0  0  | c = P
 1  0  1 79 24  1  0  | d = M
 6  1  1 37 36  0  0  | e = R
 0  0  0  4 13  1  0  | f = C
 0  0  0  0  0  0  0  | g = S
    
```

**Figura 13: Resultados del Experimento 2, obtenidos por WEKA.**

Se puede apreciar que en el caso de la *Estructura (S)*, se tiene como  $F-Measure=0$ , ya que se decidió no tomarlo en cuenta. En cuanto a las demás categorías, su  $F-Measure$  mejoraron un poco respecto al Experimento 1. Algo que se puede resaltar de esta evaluación es el porcentaje de oraciones clasificadas correctamente por el clasificador ascendió a 44.26%.

Teniendo como base los resultados de estos experimentos, se espera que con la implementación de los extractores de los atributos **Expresión** e **Histórico**, así como la reducción de las oraciones categorizadas como *Estructura (S)*, la cantidad de oraciones clasificadas correctamente por el clasificador incremente significativamente.

En el Experimento 3, ya se tiene incluido el atributo **Expresión**, mas no el **Histórico**. Se pudo ver que la precisión del clasificador mejoró bastante en comparación de los experimentos anteriores. Los resultados de este experimento se pueden apreciar en la **Figura 14**.



```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      299          65.4267 %
Incorrectly Classified Instances    158          34.5733 %
Kappa statistic                    0.5723
Mean absolute error                 0.1429
Root mean squared error             0.2622
Relative absolute error             60.2971 %
Root relative squared error         76.1933 %
Total Number of Instances          457

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
          0.625   0.031   0.769     0.625   0.69       0.881    B
          0.55    0.026   0.667     0.55    0.603     0.854    G
          0.639   0.01    0.907     0.639   0.75      0.878    P
          0.803   0.276   0.5       0.803   0.616     0.84     M
          0.487   0.095   0.514     0.487   0.5       0.827    R
          0.222   0.002   0.8       0.222   0.348     0.906    C
          0.785   0       1         0.785   0.879     0.935    S
Weighted Avg.  0.654   0.095   0.707     0.654   0.658     0.869

=== Confusion Matrix ===

 a  b  c  d  e  f  g  <-- classified as
40  2  0  18  4  0  0  | a = B
 4  22  1  9  4  0  0  | b = G
 4  3  39  11  4  0  0  | c = P
 1  4  0  94  17  1  0  | d = M
 3  1  1  35  38  0  0  | e = R
 0  1  2  7  4  4  0  | f = C
 0  0  0  14  3  0  62  | g = S
    
```

**Figura 14: Resultados del Experimento 3, obtenidos por WEKA.**

Se puede ver que el *F-Measure* de todas las categorías ha aumentado considerablemente, incluso el de la categoría Brecha ha dejado de ser 0. Esto se debe a que el atributo *Expresión* ha hecho que el modelo clasificador haga una diferenciación más clara entre las diferentes categorías. Como se puede ver en el matriz de confusión, la cantidad de aciertos en la clasificación de las categorías ha mejorado bastante, teniendo así un 65.4% de oraciones clasificadas correctamente por el clasificador.

Cabe resaltar que en este último experimento se volvió a tomar en cuenta la categoría *Estructura* (S), pues al hacer la prueba sin esta categoría, la precisión del clasificador decayó a 61%. Así también, no se tomó en cuenta el atributo **Modal**, pues la precisión decaía a 64.9%.

### 3 Consideraciones Finales

Teniendo como base los resultados de los experimentos anteriores, se puede concluir que el clasificador AZEsp tiene una precisión de 65%, el cual es suficiente para utilizarlo en el ambiente SciEsp. Por ese motivo, y la limitación del tiempo del proyecto, se decidió no implementar el atributo **Histórico**. Por otro lado, cabe resaltar que el desempeño del clasificador aumentaría generalizando la lista de expresiones comunes (ver **Tabla 17**), utilizando expresiones regulares<sup>3</sup>.

<sup>3</sup> Una expresión regular, a menudo llamada también regex, es una secuencia de caracteres que forma un patrón de búsqueda, principalmente utilizada para la búsqueda de patrones de cadenas de caracteres u operaciones de sustituciones (Wikipedia, 2013).

## CAPÍTULO 7: Conclusiones y trabajos futuros

Se sabe que la redacción del resumen de los textos científicos es una técnica básica y fundamental para la organización de ideas y preparación de información para redactar correctamente textos científicos más complejos. Por tal motivo, el presente proyecto de fin de carrera presenta la implementación de un software de apoyo a la escritura de resúmenes de textos científicos en español, el cual ayudará al escritor a redactar resúmenes de sus textos científicos con una estructura adecuada.

Antes de empezar con el desarrollo del ambiente de ayuda a la redacción de textos científicos en español, se investigó si éste era viable. Se constató la existencia de un ambiente de ayuda para el idioma portugués, y se analizó sus características. Como resultado de este análisis, se consideró posible la creación de un ambiente para el español. Cabe resaltar que no se encontró ninguna iniciativa para esta lengua.

Este trabajo contribuye con un sistema de categorización retórica de resúmenes de textos científicos en español, que puede servir de punto de partida para otros proyectos del mismo rubro con un alcance más profundo. Por ejemplo, otros proyectos podrían abarcar, no sólo resúmenes, sino también introducciones, metodologías, discusiones, conclusiones, entre otros; incluso podría ayudar a proyectos donde se desea implementar la herramienta en otro idioma.

El sistema, denominado AZEsp, tiene como base la técnica *Argumentative Zoning* (FELTRIM et al., 2001) y fue adaptado para clasificar las oraciones de resúmenes de textos científicos en español en una de las 7 categorías retóricas definidas en el proyecto SciEsp: Contexto, Brecha, Propósito, Metodología, Resultado, Conclusión y Estructura. La precisión de AZEsp (65.4%), está un poco por debajo de la obtenida por AZPort (75%), perteneciente al proyecto SciPo (FELTRIM et al. 2003). Sin embargo, los resultados de los experimentos realizados para AZEsp, como parte del ambiente SciEsp, demuestran que la precisión actual (65.4%) es suficiente para su utilización. Es decir, debido a la buena eficiencia de AZEsp, los estudiantes universitarios podrán emplear esta herramienta para la redacción de sus resúmenes; ellos podrán identificar sus errores y deficiencias en la redacción, y serán capaces de mejorar de forma autodidacta.

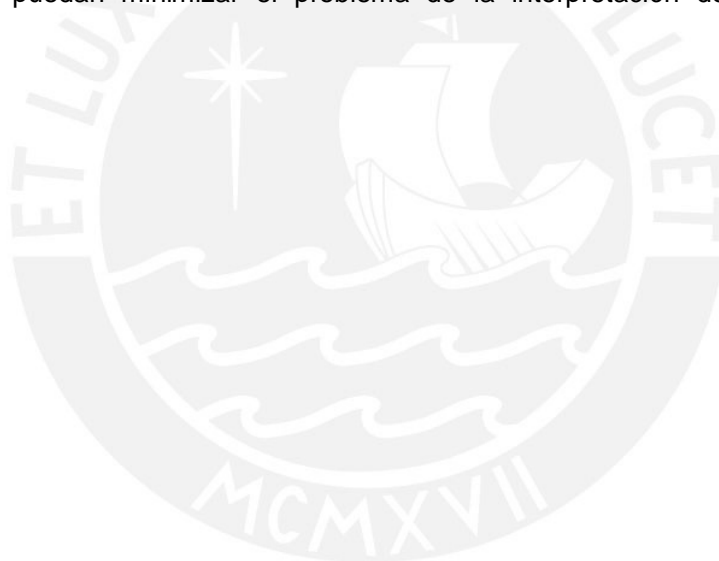
En el desarrollo de AZEsp, fueron implementados algoritmos para la extracción automática de los valores de los atributos de clasificación. Uno de estos algoritmos fue desarrollado para la extracción de características sintácticas (Tiempo, Voz y Modal) de los verbos, donde se tuvo que emplear la librería Freeling y se tuvo que codificar de tal forma que sea aplicable a la lengua española. A pesar de que este algoritmo sea un producto intermedio de AZEsp, no deja de ser una contribución de este proyecto de fin de carrera, ya que puede ser utilizada por cualquier aplicación que necesite extraer dichos atributos de los verbos en español. Además, este algoritmo se puede emplear como modelo para extraer otras características sintácticas de los verbos.

Otra contribución de este trabajo son los resultados del análisis del corpus de resúmenes de textos científicos del área de Informática. Fueron observados el estilo y la forma corriente de escribir estos resúmenes, así como los típicos errores, recurrentes de la comunidad académica analizada, tales como los alumnos de pregrado de la facultad de Informática de la PUCP (30 resúmenes), y estudiantes de posgrado de la facultad de Informática de diversas universidades extranjeras (14 resúmenes). Por otro lado, los resultados estadísticos obtenidos del análisis del corpus ayudan a apreciar claramente la situación actual de los alumnos de la PUCP en comparación con las demás universidades, en cuanto a la redacción de este tipo de textos, lo que nos permite tomar decisiones para mejorar esta situación.

Como extensión inmediata de este trabajo, se pretende tratar “introducciones de textos científicos en español”, utilizando una metodología similar a la empleada para los resúmenes. Esto incluye, además del análisis de un corpus de introducciones, la extensión del sistema de detección automática de estructura esquemática AZEsp, el cual actualmente está implementado solo para el análisis de resúmenes. Tal tarea implica no solo la adecuación de los recursos utilizados en el proceso de extracción de los atributos ya implementados, sino también la implementación y análisis de impacto de otros posibles atributos. Por otro lado, sería interesante verificar el desempeño de dicho sistema utilizando un modelo clasificador, diferente al de Naive Bayes.

Como se mencionó anteriormente, se tiene la intención de expandir las funcionalidades del ambiente SciEsp. Como trabajo futuro, también se espera trabajar en el desarrollo de una sección dedicada a la evaluación de los textos académicos, no solo en los aspectos estructurales ya cubiertos por SciEsp, sino también en otros aspectos que también influyen en la evaluación de un texto, como la coherencia, la cohesión y estilo.

Otro de los temas que se abordarán en el trabajo futuro es la evaluación de la interfaz del prototipo SciEsp, en base a criterios de usabilidad, con el objetivo de implementar mejoras que puedan minimizar el problema de la interpretación de los modelos estructurales.



## Referencias bibliográficas

- ALLEN, J.  
1995 Natural Language Understanding. Redwood City: Benjamin/Cummings.
- ANTIQUERA, L., FELTRIM, V.D. & NUNES, M.G.V.  
2003 Projeto e Implementação do Sistema SciPo. Relatórios Técnicos do ICMC-USP, n. 223, Dezembro, 2003, São Carlos, pp. 45.
- ATTALI, Y. & BURSTEIN, J.  
2006 "Automated Essay Scoring with e-rater V.2". Journal of Technology, Learning, and Assessment (JTLA), Volumen 4, N°3.
- BETSY  
s/f Consulta: 20 de abril del 2013  
< <http://echo.edres.org:8080/betsy> >
- BURSTEIN, J.  
2003 The e-rater scoring engine: Automated Essay Scoring with natural language processing. In M. D. Shermis and J. C. Burstein (Eds.), *Automated Essay Scoring: A cross disciplinary approach* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates.
- CASELI, H.M., FELTRIM, V.D. & NUNES, M.G.V.  
2002 TagAlign : uma ferramenta de pré-processamento de textos. Relatórios Técnicos do ICMC-USP, n. 169, Junho, 2002, São Carlos, pp. 38.
- COMPARÁN, J. J. et al.  
2007 Lengua Española 3. México: Umbral Editorial.
- CONTRERAS, H.  
2001 Procesamiento del Lenguaje Natural basado en una "gramática de estilos" para el idioma español. Tesis presentada para obtener el título de Licenciatura en Ingeniería Informática. Bogotá: Universidad de los Andes.
- COVINGTON, M.  
1994 "Natural Language Processing for Prolog Programmers". Artificial Intelligence Programs The University of Georgia Athens, Georgia: PRENTICE HALL, Englewood Cliffs.
- DIAS, B. C.  
1996 A FACE TECNOLÓGICA DOS ESTUDOS DA LINGUAGEM: o processamento automático das línguas naturais. Tese apresentada para obtenção do Título de DOUTOR em LETRAS – na área de concentração Lingüística e Língua Portuguesa – à Faculdade de Ciências e Letras da Universidade Estadual Paulista.
- Diccionario de la Real Academia Española (RAE)  
2012 "Corpus". Madrid. Consulta: 09 de junio del 2013.  
< <http://lema.rae.es/drae/?val=corpus> >
- Diccionario de la Real Academia Española (RAE)  
2012 "Redactar". Madrid. Consulta: 21 de abril del 2013.  
< <http://lema.rae.es/drae/?val=Redactar> >

ELLIOT, S.

2003 IntelliMetric: from here to validity. In Mark D. Shermis and Jill C. Burstein (Eds.). Automated essay scoring: a cross disciplinary approach. Mahwah, NJ: Lawrence Erlbaum Associates.

Equipo de trabajo del curso de Redacción y Comunicación de EE.GG.CC de la PUCP  
2008 Redacción y Comunicación – Material de trabajo para el alumno. Lima: PUCP.

Estudios Generales de Ciencias de la PUCP

s/f Consulta: 20 de abril del 2013

< <http://facultad.pucp.edu.pe/generales-ciencias> >

FELTRIM, V.; ALUÍSIO, S.; NUNES, M.

2003 Analysis of the rhetorical structure of computer science abstracts in Portuguese. In Proceedings of Corpus Linguistics, p. 212-218.

FELTRIM, V.D, NUNES, M.G.V. & ALUISIO S.M.

2001 Um Corpus de Textos Científicos em Português para a Análise da Estrutura Esquemática. Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil.

FELTRIM, V.D., TEUFEL, S., NUNES, M.G.V. & ALUÍSIO, S.M.

s/f Argumentative Zoning Applied to Critiquing Novices' Scientific Abstracts. In Yan Qu, James G. Shanahan and Janyce Wiebe (eds.) Exploring Attitude and Affect in Text: Theories and Applications.

FOLTZ, P. W.

1996 "Latent Semantic Analysis for text-based research". Behavior Research Methods, Instruments and Computers, Volumen 28, N° 2, pp. 197–202.

FREELING

s/f Consultado 09 de noviembre del 2013.

< <http://nlp.lsi.upc.edu/freeling/> >

GRISHMAN, R.

1986 Computational Linguistics: an introduction. Cambridge: Cambridge University Press.

HIRSH, N. & LIMO, A.

2006 Consecuencias sociales del contacto lingüístico: Diglosia y actitudes lingüísticas. ¿Cambio o muerte de las lenguas?: Reflexiones sobre la diversidad lingüística, social y cultural del Perú, Lima: UPC.

HUAMÁN, M. A.

s/f Cómo escribir un artículo científico. Boletín 44. Universidad Nacional Mayor de San Marcos.

Instituto Nacional de Estadística e Informática (INEI)

2013 Consulta: 15 de abril del 2013.

< <http://www.inei.gob.pe> >

- LANDAUER, T. K., LAHAM, D., & FOLTZ, P. W.  
2003 Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor. In M. Shermis & J. Bernstein, (Eds.). Automated Essay Scoring: A cross-disciplinary perspective. Mahwah, NJ: Lawrence Erlbaum Publishers.
- MANNING, C,  
2000 Foundations of Statistical Natural Language Processing. Segunda edición. Londres, Inglaterra: Instituto tecnológico de Massachusetts.
- Ministerio de Educación  
2010 "Indicadores". Consulta: 15 de abril del 2013.  
< <http://escale.minedu.gob.pe/indicadores> >
- MANARIS B. Z. & SLATOR B. M.  
1996 Interactive Natural Language Processing: Building on Success, Computer, IEEE.
- MOLESTINA, C. J. et al.  
1988 Fundamentos de comunicación científica y redacción técnica: una recopilación. Primera edición. San José, Costa Rica: Instituto Interamericano de Cooperación para la Agricultura.
- MOLLIERO, S.  
s/f Inteligencias sintéticas: un acercamiento al fascinante mundo de las máquina inteligentes. Editorial Alsina.
- MONK  
s/f Consulta: 10 de mayo del 2013.  
<<http://monkpublic.library.illinois.edu/monkmiddleware/public/tutorial/index.html>>  
>
- MORENO, A.  
1998 Lingüística Computacional: Introducción a los modelos simbólicos, estadísticos y biológicos. Madrid: Editorial Síntesis.
- NLTK  
s/f "Learning to classify text". Consulta: 10 de mayo del 2013.  
< <http://nltk.googlecode.com/svn/trunk/doc/book/ch06.html> >
- PADRÓ, L.  
2011 Analizadores Multilingües en FreeLing. Linguamatica, vol. 3, n. 2, pg. 13--20. December, 2011.
- PADRÓ, L. & STANILOVSKY, E.  
2012 FreeLing 3.0: Towards Wider Multilinguality. Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey. May, 2012.
- PROJECT MANAGEMENT INSTITUTE (PMI)  
2008 Project Management Body Of Knowledge (PMBOK). Cuarta Edición. Pensilvania.
- RAMOS, Moisés  
2011 "El problema de comprensión y producción de textos en el Perú". Docencia Universitaria. Lima. Año 5, N° 1.

- RUDNER, L., GARCIA, V. & WELCH, C.  
2000 "The debate on automated essay grading". Journal of Technology, Learning and Assessment. Volumen 5, N° 1.
- RUDNER, L. & LIANG, T.  
2002 "Automated essay scoring using Bayes' theorem". Journal of Technology, Learning and Assessment. Volumen 1, N° 2.
- SANCHEZ, Carlos  
2005 "Los problemas de redacción de los estudiantes costarricenses: Una propuesta de revisión desde la lingüística del texto". Filología, Lingüística y Literatura. Costa Rica, Volumen XXXI, N° 1, pp. 267-295.
- SABAJ, Omar  
2009 "Descubriendo algunos problemas en la redacción de Artículos de Investigación Científica (AIC) de alumnos de postgrado". Revista Signos. Volumen 42, N° 69, pp. 107-127.
- SciPo  
2000 Consultado 20 de abril del 2013  
< <http://www.nilc.icmc.usp.br/~scipo> >
- SciPo-Farmacia  
s/f Consultado 20 de abril del 2013  
< <http://www.nilc.icmc.usp.br/scipo-farmacia> >
- SEPULVEDA, Lianet  
2012 Escrita Científica em Português por Hispanofalantes: Recursos Linguístico-computacionais baseados em Métodos de Alinhamento de Textos Paralelos e em Córpus de Aprendizes. Tesis presentada para el examen de calificación como parte de los requisitos para obtener el título de Doctor en Ciencias - Ciencias de la Computación y Matemática Computacional. São Carlos: Instituto de Ciencias Matemáticas y Ciencias de la Computación - ICMC-USP.
- SCRUM  
2011 Consultado 05 de junio del 2013.  
< <http://www.scrum.org/> >
- SLAFER, G. A.  
2009 "¿Cómo escribir un artículo científico?". Revista de Investigación en Educación. N° 6, pp.124-132.
- SQUILLARI, R., BONO, A. & RINAUDO, M.  
2000 Tareas de aprendizaje en la universidad. Análisis de producciones monográficas de estudiantes universitarios. Ponencia presentada en las X Jornadas de Producción y Reflexión sobre Educación. Universidad Nacional de Río Cuarto.
- TEUFEL, S.; CARLETTA, J. & MOENS, M.  
1999 An annotation scheme for discourse-level argumentation in research articles. In Proceedings of the Ninth Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99), 110-117.

TEUFEL, S. & MOENS, M.

2002 Summarising Scientific Articles – Experiments with Relevance and Rhetorical Status. Computational Linguistics, 28 (4), 409-446.

UNESCO

1951 Guía para la preparación y publicación de resúmenes analíticos. París, Francia.

VALIENTE, Gabriel

1997 Composición de textos científicos con Latex. Primera edición. Barcelona: UPC.

WEKA 3: DATA MINING SOFTWARE IN JAVA

s/f Consultado 09 de noviembre del 2013.  
< <http://www.cs.waikato.ac.nz/ml/weka/> >

Wikipedia

2013 “Expresión regular”. Consulta: 10 de noviembre del 2013.

< [http://es.wikipedia.org/wiki/Expresi%C3%B3n\\_regular](http://es.wikipedia.org/wiki/Expresi%C3%B3n_regular) >

“Tiempos verbales en español”. Consulta: 10 de noviembre del 2013.

< [http://es.wikipedia.org/wiki/Tiempos\\_verbales\\_en\\_espa%C3%B1ol](http://es.wikipedia.org/wiki/Tiempos_verbales_en_espa%C3%B1ol) >

