Centro Nacional de Investigación y Desarrollo Tecnológico

Subdirección Académica

Departamento de Ciencias Computacionales

TESIS DE MAESTRÍA EN CIENCIAS

Desarrollo de un sistema de resumen y traducción multi-documentos a través de un algoritmo de resumen automático basado en entidades nombradas.

presentada por
Ing. Arturo Michel Gómez Flores

como requisito para la obtención del grado de
Maestro en Ciencias de la Computación

Director de tesis
Dr. Noé Alejandro Castro Sánchez

Codirector de tesis
Dra. Azucena Montes Rendón

Cuernavaca, Morelos, México. enero de 2018

"Año del Centenario de la Promulgación de la Constitución Política de los Estados Unidos Mexicanos"

Cuernavaca, Morelos a 09 de enero del 2018
OFICIO No. DCC/001/2018

Asunto: Aceptación de documento de tesis

DR. GERARDO V. GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial del **Ing. Arturo Michel Gómez Flores**, con número de control M15CE080, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis profesional titulado **"Desarrollo de un sistema de resumen y traducción multi-documentos a través de un algoritmo de resumen automático basado en entidades nombradas"** y hemos encontrado que se han realizado todas las correcciones y observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

DIRECTOR DE TESIS

Dr. Noé Alejandro Castro Sánchez
Doctor en Ciencias de la Computación
08701806

CO-DIRECTOR

Dra. Azucena Montes Rendón
Doctora en Ciencias
4001014

REVISOR 2

M.C. Gerardo Reyes Salgado
Maestro en Ciencias de la Computación
2493370

REVISOR 1

Dr. Juan Gabriel González Serna
Doctor en Ciencias de la Computación
7820329

C.p. M.T.I. María Elena Gómez Torres - Jefa del Departamento de Servicios Escolares.
Estudiante
Expediente

NACS/lmz

Cuernavaca, Mor., 24 de enero de 2018
OFICIO No. SAC/070/2018

Asunto: Autorización de impresión de tesis

ING. ARTURO MICHEL GÓMEZ FLORES
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS
DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Desarrollo de un sistema de resumen y traducción multi-documentos a través de un algoritmo de resumen automático basado en entidades nombradas"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

ATENTAMENTE
"CONOCIMIENTO Y TECNOLOGÍA AL SERVICIO DE MÉXICO"

DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO

SEP TecNM
CENTRO NACIONAL
DE INVESTIGACIÓN
Y DESARROLLO
TECNOLÓGICO
SUBDIRECCIÓN
ACADÉMICA

C.p. M.T.I. María Elena Gómez Torres .- Jefa del Departamento de Servicios Escolares.
Expediente

GVGR/mcr

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico
Interior Internado Palmira S/N, Col. Palmira, C.P. 62490, Cuernavaca, Mor.
Tels. (01) 777 362-77-70 Ext. 4106, e-mail: *direccion@cenidet.edu.mx*
www.cenidet.edu.mx

Dedicatoria

Este gran logro en mi vida se lo dedico principalmente a mi familia que ha sido parte de ello apoyándome con su cariño, amor y comprensión para salir a delante y que nunca me dejo solo, esto es para ustedes mamá Zita Flores Estrada, papá Patricio Arturo Gómez Sánchez y mi hermana Dulce Gabriela Gómez Flores ¡los quiero mucho!

También se lo dedico a Dios quien me a permitido llegar hasta aquí haciéndome superar grandes retos.

Este trabajo también se lo dedico a mis directores la Dra. Azucena montes Rendón, al Dr. Noé Alejandro Sánchez, al Dr. Juan Manuel Torres Moreno a mis revisores

El Dr. Juan Gabriel Gonzáles Serna y el Dr. Gerardo Reyes Salgado.

Gracias a ustedes por Guiarme con sabiduría para que este gran momento llegara.

¡MUCHAS GRACIAS A TODOS!

Agradecimientos

A Dios

Por haberme otorgado la oportunidad de concluir este logro más en mi vida bendiciéndome cada día con salud y principalmente permitirme estar unido a mi familia. ¡GRACIAS DIOS MIO!

A mi mamá Zita Flores Estrada

Gracias mamá por siempre apoyarme en todo, en mis metas y decisiones, gracias por todos tus consejos que me han ayudado a ser feliz, te agradezco por ser como eres la mamá más linda y hermosa del mundo que me da comprensión y siempre estas apoyándome con mis estudios y mis gustos y ahora estoy feliz por haber logrado terminar la maestría la cual te la dedico porque tu eres parte de este logro, quiero que sepas que siempre vas a contar conmigo. ¡GRACIAS MAMÁ! ¡TE AMO!

A mi papá Patricio Arturo Gómez Sánchez

Te agradezco papá por todo el apoyo que me brindas todos los días, por siempre aconsejarme y querer lo mejor para mí, gracias por tus buenos deseos y por comprenderme en situaciones difíciles, hoy gracias a ti también puedo terminar una etapa más en mi vida y te la dedico porque eres parte de ella, quiero que sepas que siempre vas a poder contar conmigo papá, así como yo he contado con tu apoyo incondicional. ¡GRACIAS PAPÁ! ¡TE AMO!

A mi hermana Dulce Gabriela Gómez Flores

¡Hermanita! Muchísimas gracias por todo el apoyo que me has brindado incondicionalmente y por todos tus consejos y regaños que se que lo haces porque me quieres mucho, hoy estoy concluyendo esta etapa más y quiero dedicártela también porque eres parte de esto, por el gran apoyo que me diste en mi estancia a Francia y jamás me dejaste solo, sabes que yo siempre estaré para ti cuando me necesites. ¡GRACIAS HERMANA! ¡TE AMO!

Al CONACYT

Quiero darle las gracias al Consejo Nacional de Ciencia y Tecnología (CONACYT) por brindarme la beca para realizar mis estudios ya que sin ella no hubiera podido lograrlos.

Al CENIDET

Agradezco también al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) por permitirme realizar mis estudios de maestría en ciencias y a todo su personal académico que contribuyo en este logro.

A mi amigo Manuel Flores Rosales

Amigo gracias por ser mi mejor amigo en la escuela y en mi vida, por ese apoyo y enseñanza que me brindaste para sacar adelante mis proyectos y sobre todo por tu amistad y aunque vivamos lejos sabes que cuentas conmigo siempre amigo. ¡ya tu sabe!

A mi amiga Karen Jannete Jaime Díaz

Amiga muchas gracias por esa amistad que brindaste en el trascurso de la maestría y que es para siempre esta bonita amistad, por tus consejos y apoyo dentro y fuera de la escuela, sabes que cuentas siempre conmigo, te quiero ¡muchas gracias jovena!

A la Dra. Azucena Montes Rendón

Doctora Azucena le agradezco mucho el permitirme trabajar con usted el tema de tesis y apoyarme con sus correcciones y el gran ejemplo que me dio, también quiero agradecerle esa grandísima oportunidad que me brindo de poder realizar una estancia en el extranjero, es algo que siempre voy a recordar muchísimas gracias por todo Dra. Azucena. ¡MUCHAS GRACIAS!

Al Dr. Juan Manuel Torres Moreno

Doctor Juan Manuel le agradezco de manera infinita todo el apoyo que me brindó en mi estancia en Francia, usted ha sido pieza clave para sacar adelante mi tema de tesis y es por eso que le quiero dar las gracias porque me guio correctamente en el aprendizaje de nuevas herramientas que me sirvieron de mucho en mi trabajo y no solamente en eso también le agradezco por el apoyo en todo lo que necesité en la estancia independientemente de la escuela. ¡MUCHAS GRACIAS!

A la Dra. Alicia Martínez Rebollar

Doctora Alicia quiero darle las gracias por aceptarme en el CENIDET ya que usted me realizó la entrevista y hoy le demuestro que cumplí con lo prometido terminando mi maestría satisfactoriamente, muchas gracias por brindarme esta oportunidad que no desaproveché en ningún momento. ¡MUCHAS GRACIAS!

Al Dr. Noé Alejandro Castro Sánchez

Doctor Noé le quiero agradecer el asesorarme en el camino de mi maestría con sus conocimientos dándome nuevas ideas y correcciones para la mejora del mismo, le agradezco su tiempo, dedicación y apoyo para poder lograr este momento. ¡MUCHAS GRACIAS!

Al Dr. Juan Gabriel Gonzáles Serna

Doctor Juan Gabriel quiero agradecerle su presencia en mis presentaciones y cada una de sus sugerencias y opiniones que me realizó para poder mejorar mi trabajo, también le agradezco su tiempo y dedicación que me brindó. ¡MUCHAS GRACIAS!

Al Dr. Gerardo Reyes Salgado

Doctor Gerardo Reyes le doy las gracias por brindarme su tiempo, dedicación y correcciones para mi tesis ya que han sido de gran ayuda para lograr terminar mi trabajo. ¡MUCHAS GRACIAS!

A la Maestra Patricia Armas León

Teacher Pati le quiero agradecer muchísimo todo el apoyo que me ha brindado tanto en su materia como en la amistad, es una gran maestra y me da mucho gusto haberla conocido, en verdad por todo ¡MUCHAS GRACIAS!

A mis compañeros

También quiero agradecer el compañerismo y amistad de todos ustedes Manuel flores, Karen Jannete, Alberto Iturbe, Leonel Gonzales, Luis Moreno, Fernando Patiño, Hermilo Benito, Joshua Ono, Crhistian Becerra, Monse Rojas, Alondra Nava, Itzel y Kari Gama.

Resumen

En esta tesis se creó una herramienta que permite procesar una gran cantidad de documentos para generar un resumen automático basado en entidades nombradas y traducirlo a otro idioma, este idioma puede ser en español, inglés o francés.

Este trabajo se inició con el estudio del estado del arte de sistemas para la generación de resumen automático, con ello se identificó el que presenta mejores resultados y se utilizó como referencia para la evaluación de 3 algoritmos creados en esta tesis.

Estos algoritmos generan resumen basándose en entidades nombradas, el primero se basa en nombres propios (NP), el segundo en verbos (V) y el tercero en ambos, dichas evaluaciones se realizaron mediante la métrica Fresa y con un corpus conformado por 500 documentos en formato de texto plano, también se realizaron pruebas con textos cortos de 50, 40 y 30 líneas. Una vez que se obtuvieron los 3 mejores algoritmos, se realizó una combinación de ellos obteniendo un algoritmo híbrido. Se realizó una evaluación final mediante la cual se determinó el mejor sistema de resumen para ser utilizado en la herramienta creada en esta tesis, el cual fue Stonv (*Summary Text based on prOper Nouns and Verbs).* Los resultados fueron los siguientes:
Stonv 0.40874, Híbrido 0.40313, Nasve 0.39328, Artex 0.38298 y Nason 0.33042.

Por otro lado, se investigaron sistemas de traducción automática y se evaluaron mediante la métrica Fresa, el sistema de traducción mejor evaluado se utilizó para combinarlo con el algoritmo Stonv. Finalmente se realizaron evaluaciones a la combinación de estos dos sistemas para obtener la forma más conveniente de unirlos y así implementar un sistema web como una sola herramienta que realiza resumen y traducción teniendo 3 idiomas como alternativas a elegir, resultando ser un sistema multi-documentos y multi-idiomas. El objetivo de evaluar resumen y traducción es el de obtener un texto con la menor cantidad de errores posibles.

Abstract

In this thesis, a tool was created to process a grand number of documents in order to generate an automatic summary based on named entities and to translate it into another language: Spanish, English or French.

This research began with the state of the art of systems'study for the generation of automatic summary, with that, it was identified the one that presents the best results, and it was used as a reference for the evaluation of 3 algorithms created in this thesis. These algorithms generate a summary based on named entities: the first is based on proper names (NP), the second on verbs (V) and the third on both. These evaluations were made using the *Fresa* metric and with a corpus made up of 500 documents in flat text format. In addition, these tests were carried out with short texts of 50, 40 and 30 lines. Once the 3 best algorithms were obtained, a combination of them was made to get a hybrid algorithm. A final evaluation was developed in which the best summary system was determined to be used in the tool created in this thesis: Stonv (Summary Text based on prOper Nouns and Verbs). The results were the following:

Stonv 0.40874, Híbrido 0.40313, Nasve 0.39328, Artex 0.38298 y Nason 0.33042.

On the other hand, several automatic translation systems were investigated and evaluated with the use of the *Fresa* metric. The best evaluated translation system was used to combine it with the Stonv algorithm. Finally, these evaluations were made for the combination of these two systems to obtain the most convenient way to join them, and thus, as a means to implement a web system as a single tool that performs summary and translation with 3 languages options. And as a result, we have a multi-document and multi-languages system.

The objective summary and translation's evaluation is to get a text with the less number of possible errors.

Índice

Índice de Figuras

Índice de Tablas

Índice de Gráficas

Capítulo 1
Introducción

1.1 Introducción

La Inteligencia Artificial (IA) se caracteriza por ser un área multidisciplinaria. Esta rama a lo largo del tiempo ha logrado avances importantes en áreas relacionadas con el ser humano creando diseños de sistemas capaces de resolver problemas utilizando como paradigma la inteligencia humana. Una de estas disciplinas es el Procesamiento de Lenguaje Natural (PLN) o por sus siglas en inglés *Natural Language Processing* (NLP). El PLN es considerado un campo interdisciplinario; se enfoca en el razonamiento mediante la lógica del pensamiento y utiliza diversas técnicas de IA. Tiene como tareas principales la traducción automática, resumen automático, búsquedas semánticas, extracción y recuperación de información, etc. El Procesamiento de Lenguaje Natural tiene como objetivo la construcción de sistemas y mecanismos capaces de emular actividades del lenguaje natural considerando principalmente aspectos lingüísticos.

En los últimos años, las tareas de traducción automática y resumen automático de textos mono-documento han avanzado en la mejora de sus técnicas. Por un lado, los sistemas que realizan resumen automático tienen por entrada un texto y como resultado un texto corto obtenido mediante técnicas basadas en bígamas o trigramas según el algoritmo.[1] Por otro lado, los sistemas de traducción automática utilizan técnicas de alineamiento y tienen conocimiento de la gramática, sintaxis, semántica etc. de ambos idiomas (origen y destino).[2]

Sin embargo, cuando se quiere realizar un resumen automático y posteriormente realizar una traducción o viceversa, los resultados no son muy satisfactorios debido a que los errores de la primera tarea se transfieren a la segunda. Esto se complica aún más cuando se trata de multi-documentos.

En la actualidad existen distintos algoritmos de resumen automáticos de textos, los cuales se distinguen por las condiciones de cada uno al momento de realizar resúmenes (extractivos o abstractivos), sin embargo, existen distintos factores que dificultan la evaluación de los resúmenes ya que no existe un único resumen valido para un texto, el lenguaje natural producido por una máquina no es 100% libre de ambigüedades, personas juzgando incrementa el coste, etc.

En este trabajo se pretende realizar un método que contemple ambas técnicas del resumen y traducción automática aplicada a multi-documento.

1.2 Objetivos

A continuación, se presenta el objetivo general de este trabajo de investigación. Posteriormente se detallan los objetivos específicos que se realizaron para concluir exitosamente la implementación de resumen y traducción automática.

1.2.1 Objetivo general

Desarrollar una aplicación que combine un algoritmo de resumen y de traducción automáticos siendo estos de los mejores en su área para resumir y traducir textos de diferentes idiomas.

1.2.2 Objetivos Específicos

- Identificar los algoritmos de resumen y traducción automáticos que procesen multi-documentos.
- Seleccionar aquellos que arrojan resultados más altos de precisión y cobertura.
- Determinar el tipo de resumen (abstractivo o extractivo) que se usará en este trabajo de investigación.
- Utilizar el traductor automático que genere mejores resultados en sus traducciones.

1.3 Justificación

En la actualidad no existe un sistema que combine resumen y traducción automática de texto que pueda tener como entrada multi-documentos y que al mismo tiempo pueda admitir texto multilenguaje, es decir, ingresar documentos escritos en más de un idioma diferente. Por otro lado, la calidad del texto que genera un sistema de resumen y traducción muchas veces no es la mejor que nos entregan. Existen distintos algoritmos de resumen automáticos de textos, los cuales se distinguen por las condiciones de cada uno al momento de realizar resúmenes (extractivos o abstractivos), sin embargo, hay distintos factores que dificultan la evaluación de los resúmenes ya que no existe un único resumen valido para un texto, el lenguaje natural producido por una máquina no es 100% libre de ambigüedades, personas corrigiendo resúmenes incrementa el costo, etc. Sin embargo, la calidad en esta tesis se mide a partir de herramientas de evaluación, así como la calidad de traductores automáticos de texto para combinarlos en una sola herramienta.

1.4 Estructura del documento

La organización de este documento se divide en siete capítulos, los cuales describen de principio a fin las etapas de desarrollo de la investigación. Con esta sección se concluye el primer capítulo, los seis restantes se puntualizan a continuación:

Capítulo II

Marco Teórico, en donde se abordan los conceptos más significativos para contextualizar al lector.

Capítulo III

Estado del Arte, en el cual se proporciona información de los trabajos de investigación que se han realizado con relación al tema de investigación de esta tesis.

Capítulo IV

Metodología de solución, en esta sección, se define claramente cada una de las fases que componen el método de solución propuesto para el problema descrito anteriormente.

Capítulo V

Diseño e Implementación, presenta detalladamente la creación de los algoritmos de resumen automático de texto basados en entidades nombradas, así como la evaluación de los mismos, la evaluación a los sistemas de traducción automática y el esquema de combinación de resumen y traducción automáticos.

Capítulo VI

Pruebas y Resultados, donde se analizan los resultados obtenidos de las distintas pruebas experimentales que se realizaron durante el trabajo de investigación.

Capítulo VII

Conclusiones, en donde se especifican los aportes y contribuciones de esta investigación, así como también los trabajos futuros relacionados con esta tesis.

.

Capítulo 2
Marco Teórico

En este capítulo se presentan los fundamentos teóricos que se revisaron previamente para el desarrollo y construcción de la metodología de esta tesis.

2.1 Procesamiento del lenguaje natural (PLN)

El procesamiento del lenguaje natural, o PLN, es una disciplina que mezcla dos perfiles profesionales para llevar a cabo diversas tareas con datos de texto o voz: por un lado, la inteligencia artificial como rama de la ingeniería en computación y, por el otro, la lingüística como rama del estudio de la lengua.[3]

El procesamiento de lenguaje natural estudia las interacciones entre las computadoras y el lenguaje humano.[4]

2.2 Resumen automático

Es una técnica de estudio que consiste en reducir un texto, expresando con las mismas palabras del autor las ideas principales vinculándolas unas con otras, sin perder la claridad expositiva. Las ideas secundarias serán incorporadas en tanto sean absolutamente necesarias para la coherencia del texto. Si se introdujeran palabras propias de quien realiza el resumen o apreciaciones personales, o se variara la secuencia, ya no sería un resumen sino una síntesis.[5]

2.2.1 Dificultad de la evaluación de un resumen

• No existe un único resumen válido para un texto.

Pueden existir más de un resumen diferente realizado al mismo texto.

• Lenguaje natural producido por una máquina.

No es 100% libre de ambigüedades el texto producido por una máquina.

• Personas juzgando incrementa el costo.

El costo de un resumen puede aumentar si se paga a personas para su corrección.

• Resumir conlleva compresión (reducción de tamaño).

Es necesario evaluar resúmenes de distintos tamaños.

• Legibilidad (puede no tener relación con la calidad del resumen).

Incoherencias generadas en el texto.

2.2.2 Tipos de evaluación

Evaluación intrínseca

El enfoque intrínseco intenta evaluar el resumen sin considerar la audiencia a la que este va dirigida. Para lograr este objetivo, se intenta dar mayor peso a aspectos como la coherencia o lo informativo del resumen generado. Es común llevar a cabo la evaluación comparando los resúmenes generados con modelos reconocidos de aquellos aspectos que un resumen de calidad deba presentar.

Estos modelos reconocidos, suelen ser resúmenes de referencia generados por sistemas de generación de resúmenes automáticos que gozan de reconocimiento, o directamente por equipos de investigación humanos.

– Calidad (por personas - No siempre acuerdo)

• Legibilidad, comprensión, acrónimos, anáforas, integridad de la estructura, gramaticalidad, estilo impersonal, etc.

– Informatividad

• Información preserva respecto al texto original (varias compresiones)

• Información contiene respeto a un resumen ideal

Evaluación extrínseca

La evaluación extrínseca de un resumen tiene su enfoque en el usuario al que va dirigido el resumen, teniendo más en cuenta la utilidad que este puede tener sobre ese usuario que su calidad como resumen. Una de las formas de evaluar ese aspecto, es asegurarse de que el texto que se usa para el resumen es fácil de entender para el usuario que lo recibe.

Evaluar el uso del resumen en otra tarea

– Encontrar documentos relevantes en una colección

– Decisión tomada leyendo el resumen o el texto original

– Sistemas de Q&A (responden a preguntas)

– Sistemas de recuperación de información

– Contenido páginas web (buscadores).[6]

2.3 Traducción automática

La Traducción Automática (TA) es traducción automatizada. Es el proceso mediante el cual se utiliza software de computadora para traducir un texto de un lenguaje natural (como el inglés) a otro (como el español).

Al procesar cualquier traducción, humana o automática, el significado del texto en el idioma original (origen) se debe restaurar totalmente en el de destino, es decir, en la traducción. Aunque en apariencia parezca sencillo, es mucho más complejo. La traducción no es una mera sustitución de una palabra por otra. Un traductor debe interpretar y analizar todos los elementos del texto y saber cómo influyen unas palabras en otras. Para ello se necesitan amplios conocimientos de gramática, sintaxis (estructura de las oraciones), semántica (significados), etc., de los idiomas de origen y de destino, además de familiaridad con cada región específica.

2.3.1 Tecnología de traducción automática basada en reglas

La traducción automática basada en reglas se basa en incontables reglas lingüísticas integradas y en millones de diccionarios bilingües para cada par de idiomas. El software analiza sintácticamente el texto y crea una representación transitoria a partir de la cual se genera el texto en el idioma de destino. Este proceso requiere léxicos amplios con información morfológica, sintáctica y semántica, además de grandes conjuntos de reglas. El software utiliza esos conjuntos de reglas complejos y, a continuación, transfiere la estructura gramatical del idioma de origen al idioma de destino.

Las traducciones se construyen con diccionarios enormes y reglas lingüísticas sofisticadas. Los usuarios pueden mejorar la calidad de la traducción instantánea añadiendo su terminología al proceso de traducción. Para ello crean diccionarios definidos por el usuario que invalidan la configuración predeterminada del sistema.

En la mayoría de los casos, hay dos pasos: una inversión inicial que aumenta de forma significativa la calidad con un costo limitado, y una inversión acumulable que aumenta la calidad de forma incremental. Aunque la TA basada en reglas proporciona a las empresas el umbral de calidad que necesitan e incluso más, el proceso de mejora de calidad puede ser largo y costoso.

2.3.2 Tecnología de traducción automática estadística

La traducción automática estadística utiliza modelos de traducción estadísticos cuyos parámetros emanan del análisis de corpus monolingües y bilingües. La creación de modelos de traducción estadísticos es un proceso rápido, pero la tecnología depende enormemente de los corpus multilingües existentes. Se necesitan un mínimo de 2 millones de palabras para un dominio específico y más incluso para el idioma en general. Teóricamente es posible alcanzar el umbral de calidad, pero la mayoría de las compañías no tienen cantidades tan grandes de corpus multilingües para crear los modelos de traducción necesarios. Además, la traducción automática estadística consume mucha CPU y requiere una configuración de hardware amplia para ejecutar los modelos de traducción que permiten obtener niveles de rendimiento promedio

2.3.3 TA basada en reglas frente a TA estadística

La TA basada en reglas proporciona una buena calidad fuera del dominio o ámbito concreto y es previsible por naturaleza. La personalización basada en diccionarios garantiza una calidad mejorada y la conformidad con la terminología corporativa. Pero a los resultados de la traducción les puede faltar la fluidez que esperan los lectores. En términos de inversión, el ciclo de personalización necesario para llegar al umbral de calidad puede ser largo y costoso. El rendimiento es alto incluso con hardware estándar.

La TA estadística proporciona una buena calidad cuando se dispone de corpus grandes y cualificados. La traducción es fluida, lo que significa que se lee bien y, por lo tanto, cumple con las expectativas del usuario. Sin embargo, la traducción no es ni previsible ni coherente. El entrenamiento a partir de corpus buenos es automático y más barato. Pero el entrenamiento sobre corpus del lenguaje general, es decir, sobre textos que no son del dominio especificado, es deficiente. Además, la TA estadística requiere un hardware determinado para crear y administrar modelos de traducción grandes. [7]

2.4 Baseline o línea base

Generalmente, una línea de base puede ser un solo producto de trabajo o conjunto de productos de trabajo que puede utilizarse como una base lógica para la comparación. También puede establecerse una línea de base como base para actividades selectas posteriores cuando los productos de trabajo cumplan ciertos criterios. Tales actividades pueden ser atribuidas con aprobación formal.

A la inversa, la configuración de un proyecto incluye a menudo una o más líneas de base, el estado de la configuración y cualquier métrica recopilada. La configuración actual se refiere al estado actual, la auditoría actual y / o las métricas actuales. De manera similar, pero con menor frecuencia, una línea base puede referirse a todos los elementos asociados con un proyecto específico. Esto puede incluir todas las revisiones de todos los elementos o sólo la última revisión de todos los elementos del proyecto, dependiendo del contexto.

Una línea de base puede ser un tipo específico de línea de base, como el cuerpo de elementos en una revisión de certificación particular.

2.5 Freeling

Freeling es una librería de código abierto para el procesamiento multilingüe automático, que proporciona una amplia gama de servicios de análisis lingüístico para diversos idiomas. Freeling ofrece a los desarrolladores de aplicaciones de Procesamiento del Lenguaje Natural funciones de análisis y anotación lingüística de textos, con la consiguiente reducción del coste de construcción de dichas aplicaciones. Freeling es personalizable y ampliable, y está fuertemente orientado a aplicaciones del mundo real en términos de velocidad y robustez. Los desarrolladores pueden utilizar los recursos lingüísticos por defecto (diccionarios, lexicones, gramáticas, etc.), ampliarlos, adaptarlos a dominios particulares, o –dado que la librería es de código abierto– desarrollar otros nuevos para idiomas específicos o necesidades especiales de las aplicaciones.[8]

2.6 Fresa

FRESA es un framework como su nombre lo dice (*Framework for Evaluating Summaries Automatically*), que tiene como objetivo proporcionar una herramienta para la evaluación automática de los sistemas informáticos que pueden resumir documentos de forma automática. FRESA calcula divergencias estadísticas entre los resúmenes; estas medidas de divergencia son*: Kullback-Leibler* (KL) y *Jensen-Shannon* (JS).[9]

Capítulo 3
Estado del Arte

3.1 Resumen

En esta sección se muestran los artículos relacionados con resumen automático.

3.1.1 La compresión de frases: un recurso para la optimización de resumen automático de documentos.

En este artículo trata sobre optimizar resúmenes automáticos generados por herramientas existentes para la disminución de tiempo de lectura del mismo, mediante la compresión de frases, los documentos que se ocuparon fueron artículos médicos.

El problema principal consiste en la creación de resúmenes largos generados por herramientas de resumen automático, siento esto un problema porque los resúmenes generados por las herramientas de resumen automático no son lo suficientemente óptimos para darles una lectura corta, siendo igual o más largos a los resúmenes que los autores del documento original realizaron, sin embardo para resolver este problema se realizaron dos métodos para la compresión de frases y se aplicaron a los resúmenes generados por herramientas de resumen automático. En la primera etapa (Extracción) se ocuparon las siguientes herramientas:

Cortex: Es un sistema de resumen automático basado en el Modelo de Espacio Vectorial (VSM), este sistema maneja la extracción de mono-documento.

Enertex: Sistema basado en VSM y enfocado a redes de neuronas, este sistema también se basa en la estadística.

Disicosum: Sistema de resumen automático de documentos médicos.

Resumidor híbrido: Se compone de varios resumidores que se unen equilibradamente para formar uno solo.

Entre otros resumidores automáticos:

SWESUM, Open Text Summarizer, Pertinence Summarizer y *Word Summarizer*.

Herramientas que se ocuparon en la segunda etapa (Compresión):

Dos estrategias manuales:

1. Eliminación manual intuitiva

2. Eliminación manual basada en la RST

Cuatro estrategias automáticas:

1. Eliminación adjetival

2. Eliminación adverbial

3. Eliminación adjetival y adverbial

4. Eliminación aleatoria base line

Conclusión

En este trabajo se demostró que se puede beneficiar un resumen aplicándole la compresión de frases, sin embargo, este método no beneficia la evaluación ROUGE ya que algunas co-ocurrencias que forman parte de esta evaluación se eliminan al momento de aplicar la compresión.

3.1.2 Un algoritmo lingüístico-estadístico para resumen automático de textos especializados

Existen algoritmos de resumen automáticos basados en sistemas estadísticos o en sistemas lingüísticos, en este artículo se maneja especialmente un algoritmo híbrido de resumen automático, el cual combina diferentes algoritmos que manejan los dos tipos de sistemas para crear uno solo algoritmo equilibrado el problema surge por la generación de resumen automático de textos largos de 4 a 5 páginas y de esto surgió la necesidad de obtener un algoritmo híbrido que resuma textos cortos y específicos de una página, sin embargo ya se cuenta con un algoritmo hibrido que se había creado anteriormente y simplemente se refino para obtener mejores resultados. Este algoritmo hibrido está compuesto por los siguientes algoritmos de resumen:

CORTEX: Sistema estadístico

YATE: Sistema lingüístico

ENERTEX: Sistema estadístico

DISICOSUM: Sistema lingüístico

El algoritmo de decisión para la selección de las frases u oraciones del resumen ocupa 4 fases:

Fase 1. Acuerdo: si una oración del texto es seleccionada por todos los sistemas, el algoritmo la mantiene.

Fase 2. Mayoría: si una oración del texto es seleccionada por la mayoría de los sistemas, el algoritmo la mantiene.

Fase 3. Score: si una oración es seleccionada solo por uno o dos sistemas, el algoritmo elije la que tenga asignado un mayor score.

Fase 4. Score + orden de las oraciones en el texto original: si se necesita una cantidad determinada de oraciones para el resumen y varias oraciones coinciden en su score, el algoritmo prioriza la que aparece en primer lugar en el texto original.

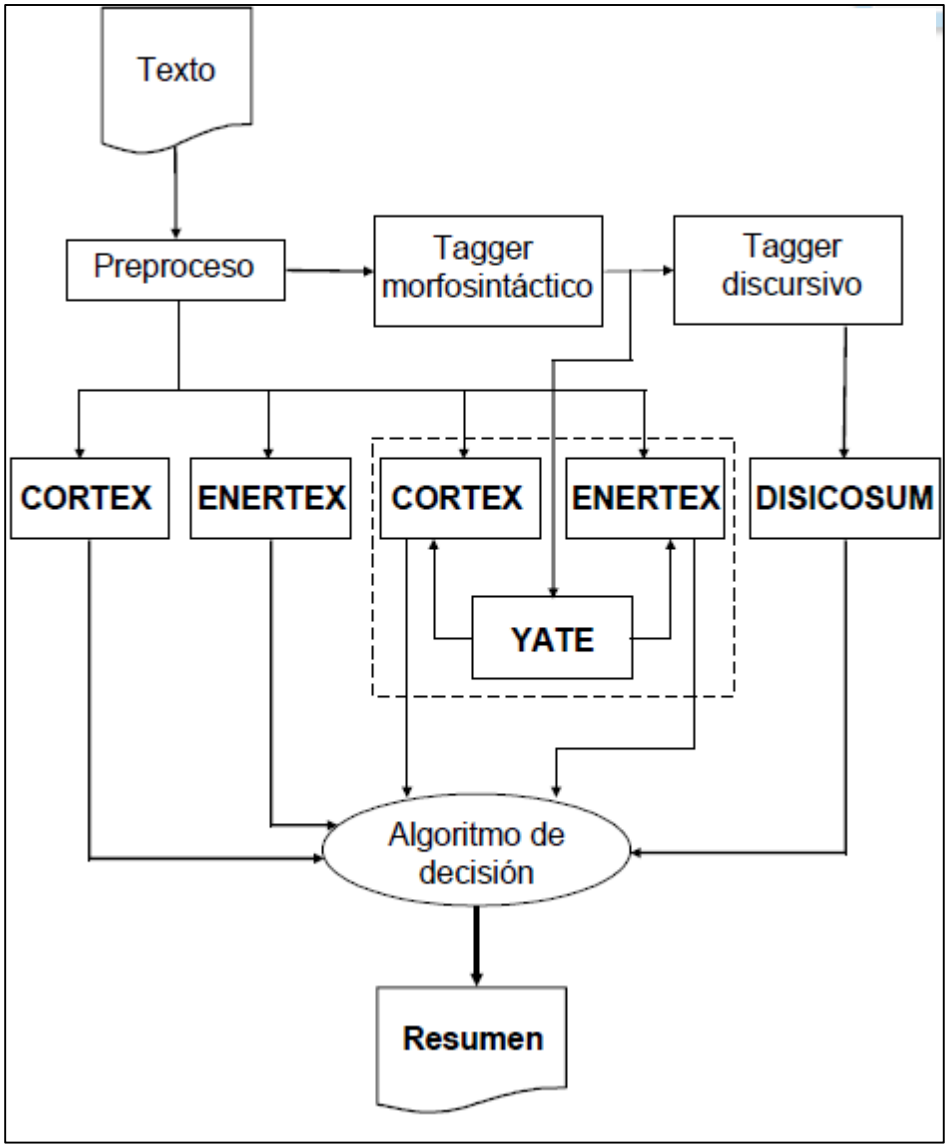


Figura 1Arquitectura del sistema de resumen híbrido lingüístico-Estadístico.

En la figura 1 se puede observar que el algoritmo híbrido está conformado por los algoritmos de resumen automático Cortex, Enertex, Yate y Disicosum, previo a estos algoritmos el texto lleva un pre-proceso antes de ser resumido por los mismos.

	ROUGE-2		ROUGE-SU4	
Sistema de resumen	**Frases**	**100 Palabras**	**Frases**	**100 P**
Híbrido	**0.3307**	**0.3163**	**0.3457**	**0.3288**
CORTEX	**0.3193**	**0.2927**	**0.3386**	0.3105
CORTEX+YATE	0.3038	0.2913	0.3281	**0.3152**
ENERTEX	0.2552	0.2314	0.2830	0.2589
DISICOSUM	0.2851	0.2750	0.3053	0.2945
$Baseline_1$	0.1454	0.1428	0.1835	0.1808
$Baseline_2$	0.1931	0.1861	0.2333	0.2260
Word	0.1873	0.1857	0.2273	0.2245
Pertinence	0.1768	0.1606	0.2266	0.2095
Swesum	0.2026	0.1939	0.2382	0.2289
OTS	0.2337	0.2176	0.2675	0.2533
Médico 1	0.3329	0.3030	0.3414	0.3130
Médico 2	0.3230	0.2993	0.3374	0.3130
Médico 3	0.3099	0.2721	0.3201	0.2898

Figura 2. Comparación de los valores de Rouge entre los diferentes resúmenes.

En la figura 2 se aprecia los resultados obtenidos con la métrica de evaluación Rouge, en la cual nos muestra que el algoritmo híbrido obtiene mejores resultados en su resumen, sin embargo, entre los algoritmos por si solos, Cortex es el algoritmo que está por encima de todos los demás.

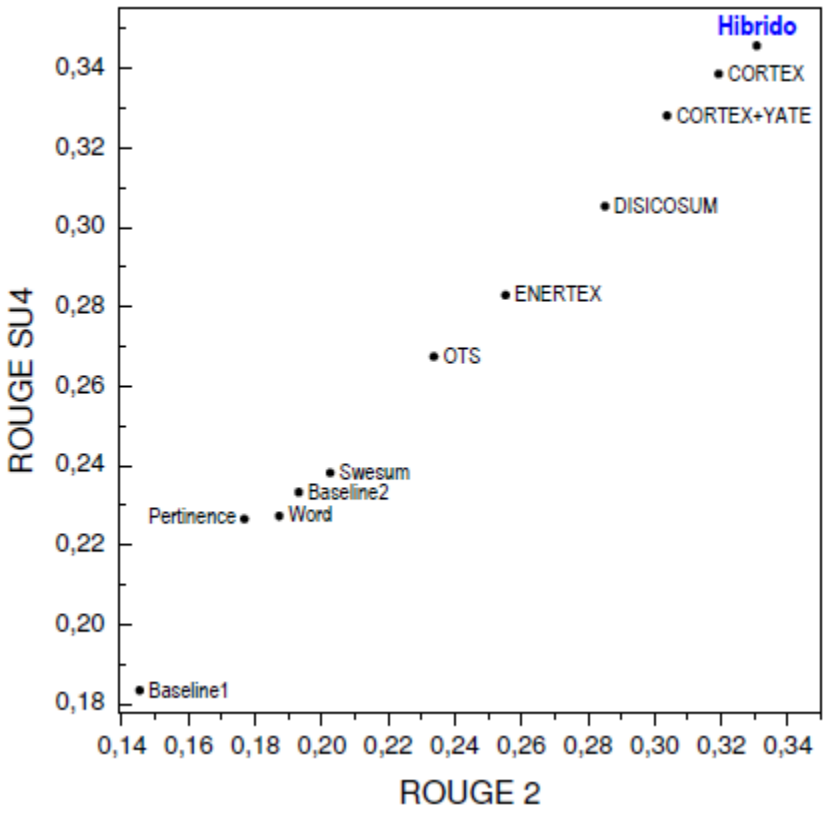


Figura 3. Evaluación ROUGE 2.

Conclusión:

Este algoritmo lingüístico-estadístico consigue superar a los otros sistemas de resumen y a las líneas de base diseñadas. Además, los resúmenes que ofrece son muy similares a los realizados por médicos especialistas, lo cual es una evidencia clara de la calidad de los mismos.

3.1.3 ARTEX is AnotheR TEXt summarizer

En este artículo se presenta un nuevo método de resumen basado en la extracción de frases. Modelo de Espacio Vectorial (VSM). Marcamos cada oración mediante el cálculo de su producto interior con un vector de pseudo-frase y un vector de pseudo-palabra. Los resultados muestran que no sólo Artex conserva el contenido de los resúmenes generada usando esta nueva representación, pero a menudo, sorprendentemente el rendimiento puede ser mejorado. Artex podría ser un algoritmo interesante y sencillamente utilizando el paradigma de integración extractiva. Nuestras pruebas sobre los corpus en tres idiomas (inglés, español y francés) evaluado por el algoritmo de Fresa (sin referencias humanos) para confirmar el buen rendimiento de Artex.

Fresa es similar a la evaluación Rouge, pero los resúmenes de referencia humana no son necesarios. Fresa calcula la divergencia de probabilidades entre el resumen candidato y el origen del documento.

El paquete Fresa calcula la divergencia entre la fuente de documento y los resúmenes.

Algoritmos de lematización

- Lemmatization
- Ultra-stemming
- Fix1

Corpus español

El español es una lengua con una variabilidad mayor de inglés. Los resultados muestran que Artex supera a Cortex y Enertex.

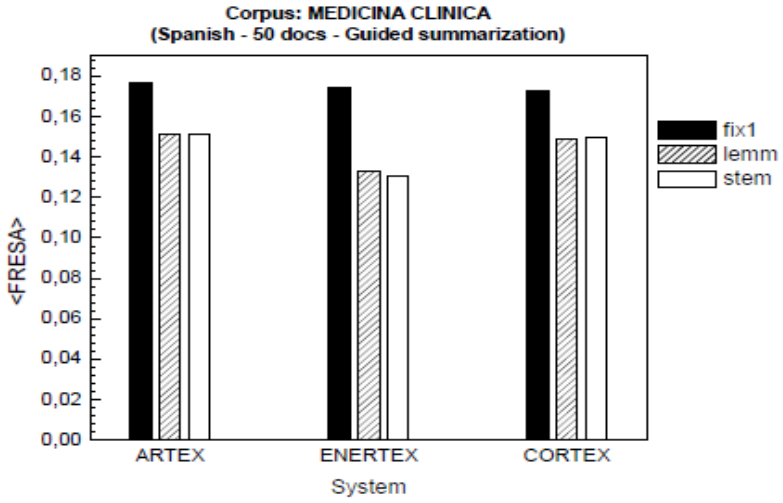


Figura 4. Grafica de evaluación Fresa del corpus en idioma español.

Corpus inglés.

En la figura 5, se muestra el desempeño de los tres resumidores utilizando *Fix1*, *stemming y lematización*. Los resultados muestran que *ultra stemming* mejora la puntuación de los tres sistemas de resumen automáticos. Artex y Cortex exponen resultados similares en el contenido de la información.

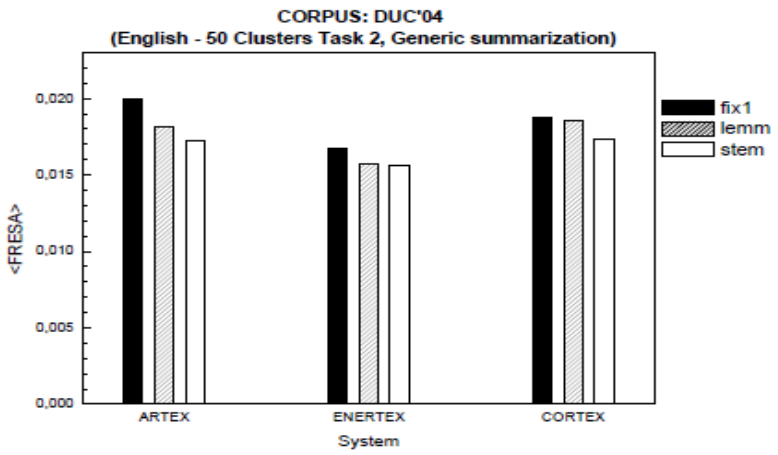


Figura 5. Grafica de evaluación Fresa del corpus en idioma inglés.

Corpus francés

El francés es una lengua con una gran variabilidad también. Enseguida se muestra la puntuación Fresa. Los resultados muestran un comportamiento similar: Ultra mejora la puntuación de los tres sistemas de resumen automático utilizados. En particular, la eficacia de Artex es menos sensible a la normalización de otros resumidores.

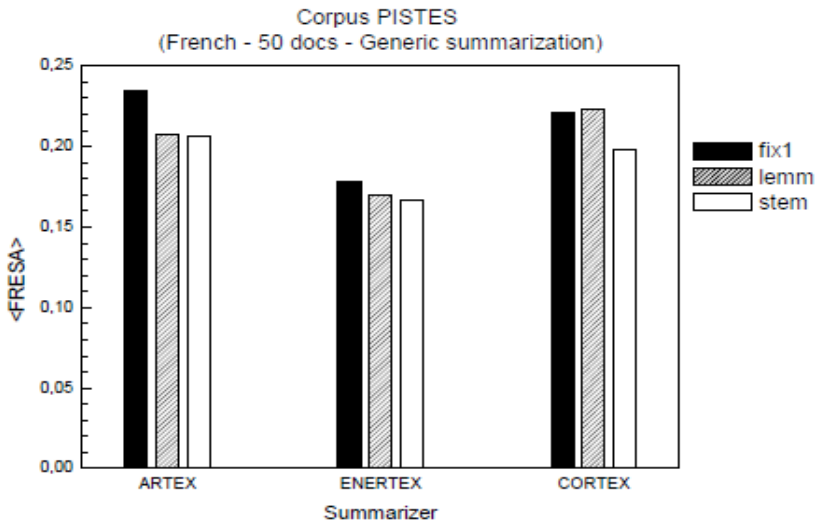


Figura 6. Grafica de evaluación Fresa del corpus en idioma francés.

La figura 6 muestra los tiempos de procesamiento para cada corpus, siguiendo el método de normalización de Cortex, Artex y Enertex. Los tiempos de procesamiento de *ultra-stemming Fix1* son más cortos en comparación con todos los métodos de otros.

Por ejemplo, Cortex es un generador de resúmenes muy rápido con O (log *p*2)

(Donde *p*P =*x* N), y los tiempos de procesamiento de *stemming y Fix1* están cerca.

Por otro lado, Enertex tiene una complejidad de O (*p* 2), entonces se necesita más tiempo para procesar el mismo cuerpo.

El rendimiento de algoritmo Artex se mantiene cerca de la Cortex.

	Summarizer Average Time (all corpora)		
Normalization	CORTEX	ARTEX	ENERTEX
Lemmatization	1.60'	2.50'	10.42'
Stemming	0.54'	1.29'	9.47'
FIX_1	**0.32'**	*0.40'*	4.25'

Figura 7. Tabla de los tiempos de procesamiento.

Conclusión

Artex es un algoritmo rápido y muy sencillo basado en el modelo VSM y el extractivo paradigma. El método utiliza una representación matricial para calcular una puntuación normalizada para cada frase, utilizando el producto interno de pseudo-(frases | palabras) vectores.

Los resultados obtenidos en inglés, francés y español muestran que Artex puede lograr buenos resultados para la calidad del contenido.

Las pruebas con otros corpus (DUC y TAC campañas de evaluación, INEX, etc.) en mono-y multi-documento guiada por un sujeto, mediante la evaluación del contenido de las evaluaciones (Rouge) o sin resúmenes de referencia todavía están en curso.

3.2 Traducción automática

En esta sección se muestran los artículos relacionados con traducción automático.

3.2.1 Estudio comparativo de traductores automáticos en línea: Systran, Reverso y Google [10]

En este artículo se realiza una evaluación de tres sistemas de traducción automática en línea: Systran, Reverso y Google, se muestran las dificultades con la que se presentan los traductores automáticos a través de la evaluación de oraciones en inglés y español, esto para determinar qué traductor obtiene mejores resultados.

El artículo nos muestra a continuación el enfoque de traducción de cada traductor.

"• Systran usa un sistema directo, es decir, para cada palabra de la lengua ~~de~~ origen hay un equivalente en la lengua meta. Los diccionarios y los mecanismos para el análisis morfológico son muy completos pero los procesos de análisis sintáctico son bastante limitados.
• Reverso utiliza sistemas de transferencia en los que podemos distinguir tres etapas en su funcionamiento. Tal como señalan Diéguez y Riedemann (1998: 214-215), en la primera fase se analiza el texto fuente en la lengua original a nivel morfológico, sintáctico y semántico. En la fase de transferencia se transforma la estructura obtenida a la que llegamos en la fase anterior a otra estructura similar en la lengua meta. En la última fase se genera la oración en la lengua meta.
• Los sistemas de traducción automática estadística, como el desarrollado por Google, generan traducciones a partir de métodos estadísticos basados en corpus de textos bilingües. Primero se alinean las oraciones de los textos de la lengua de origen y la lengua meta y después se calculan las probabilidades de traducción o de equivalencia, es decir, las probabilidades de que se correspondan con otras de traducciones realizadas por profesionales en ambas lenguas."

En este análisis se tomaron en cuenta para la evaluación las siguientes características:

- ✓ Homonimia y polisemia
- ✓ Calcos léxicos y sintácticos
- ✓ Traducción de construcciones pasivas
- ✓ Traducción de refranes
- ✓ Traducción de modismos
- ✓ Traducción adverbial
- ✓ Traducción de colocaciones
- ✓ Traducción de siglas
- ✓ Traducción nombres propios y topónimos

En anexo 1, se muestran las evaluaciones mencionadas anteriormente.

3.3 Resumen y traducción automática

En esta sesión se muestran los trabajos relacionados que utilizan traducción automática y resumen automático.

3.3.1 TESIS-Sistema resumidor-traductor automático (MOYA, 2012)

En este artículo se utilizaron dos herramientas, una para el resumen automático (ENERTEX) y la otra para la traducción automática (REVERSO) de inglés a español. Primero se traduce el texto de inglés a español y después se realiza el resumen
Se utilizó la herramienta Fresa para la evaluación de la calidad del resumen esta herramienta fue desarrollada en el Laboratorio de Informática de Aviñón (LIA) de la Universidad de Aviñón en el año 2010. Fresa es un framework, como su nombre lo dice *(Framework for Evaluating Summaries Automatically*) el cuál calcula divergencias estadísticas entre los dos resúmenes, el que fue generado por Fresa y el que generó el sistema que se desea evaluar.
La razón por la cual Fresa emplea estas divergencias es que integra un conjunto de estadísticas que permiten realizar una comparación de distribuciones de probabilidad (a diferencia de los distintos sistemas de evaluación existentes que utilizan medidas que dependen del uso de n-gramas y del procesamiento aplicado al texto de entrada como coocurrencias de n-gramas, lematización o quitar palabras de una *stop-list*). El resultado de esta comparación es un valor que puede ser usado para calificar el resumen del sistema.
Dado que el desempeño de FRESA ha sido comparado con otras herramientas como Rouge, es garantía el considerar que, si el resultado de la divergencia de los resúmenes sometidos a evaluación es más bajo que los resultados otorgados para los resúmenes que el mismo Fresa genera, el resumen evaluado es bueno y de hecho se considera mejor que el que genera Fresa.

Suplente (ROUGE (métrico)) ROUGE, o Recall Orientada para Gisting Evaluación, es un conjunto de métricas y un paquete de software que se utiliza para evaluar el resumen automático y el software de traducción automática en el procesamiento del lenguaje natural.

Reverso es un sistema de TA que funciona empleando el enfoque de TA estadístico para generar los resultados traducidos del texto fuente al idioma destino. Existe una versión gratuita en línea del sistema, la cual entrega la misma calidad que el sistema completo (cuando se compra); sólo con la limitante del tamaño del texto a traducir.
Fue desarrollado por *Softissimo*, una compañía de desarrollo de software de origen francés que desarrolla numerosas aplicaciones de TA de alto rendimiento, entre las que destacan los traductores automáticos y los diccionarios electrónicos.

Nota: La firma *Softissimo* cuenta con servicios lingüísticos profesionales que garantizan una calidad de resultados reconocida por publicaciones como: *PC expert*, *Windows News* y además ganó el *IST prize.*

Los inconvenientes con los que se topa esta tesis son los siguientes:

1. El sistema está delimitado únicamente a documentos con carácter informativo, (noticias) ya que se creó con ese fin.
2. Los documentos deben ser de texto plano, por lo cual solo admite archivos en formato txt.
3. El sistema está enfocado en traducción de idioma origen inglés al español.
4. Se observó un bajo desempeño del sistema para aquellos documentos que eran de pequeñas dimensiones (menores a 150 palabras de longitud).
5. Los resúmenes no son tan buenos cuando se realizan al 10% del documento original.

3.3.2 Columbia Newsblaster: Multilingual News Summarization on the Web

En este trabajo se realiza traducción de noticias que se encuentran en diferentes idiomas, este sistema recoge las noticias y realiza un resumen de ellas, obteniendo como salida resumen y traducción únicamente en inglés.

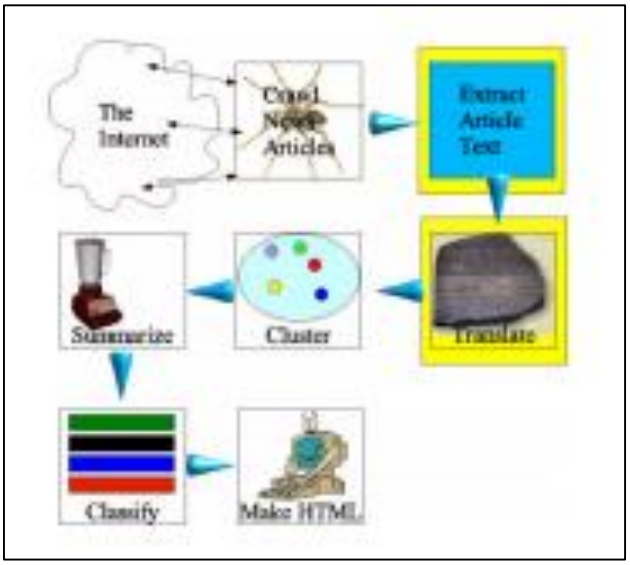


Figura 8 Metodología Newsblaster

Tabla 1 Comparación de artículos de resumen automático.

Artículo de resumen automático.	¿Cuál es el problema final?	¿Qué herramientas se utilizaron?	¿Cuáles son las desventajas?
Un algoritmo lingüístico-estadístico para resumen automático de textos especializados [11]	Se genera un resumen automático de textos largos de 4 a 5 páginas. Con un algoritmo híbrido	CORTEX: Sistema estadístico YATE: Sistema lingüístico ENERTEX: Sistema estadístico DISICOSUM: Sistema lingüístico	1. Solo se realizaron pruebas en castellano mediante artículos de medicina.
La compresión de frases: un recurso para la optimización de resumen automático de documentos. [12]	Se obtienen resúmenes largos generados por herramientas de resumen automático.	CORTEX ENERTEX DISICOSUM RESUMIDOR HIBRIDO SWESUM OPEN TEXT SUMMARIZER PERTINENCE SUMMARIZER WORD SUMMARIZER -HERRAMIENTAS PARA LA COMPRESIÓN MANUAL Y AUTOMATICA.	1. no beneficia la evaluación ROUGE ya que algunas co-ocurrencias que forman parte de esta evaluación se eliminan al momento de aplicar la compresión.
ARTEX otro resumidor de texto [1]	La calidad de los resúmenes automáticos con respecto a ARTEX	Se evaluaron los sistemas: CORTEX ARTEX ENERTEX Herramienta de evaluación: FRESA	1. El sistema CORTEX supera en tiempo a ARTEX, pero no en calidad. 2. ARTEX supero en calidad a CORTEX y a ENERTEX, aunque no fue el más rápido por una diferencia mínima. 3. ENERTEX fue el algoritmo con menor calidad que CORTEX y ARTEX.

Tabla 2 Comparación de artículos de traducción automática.

Artículos de traducción automática.	¿Cuál es el problema?	¿Qué herramientas se utilizaron?	¿Cuáles son las desventajas?
Apertium, una plataforma de código abierto para el desarrollo de sistemas de traducción automática [13]	La ambigüedad en el sentido de las palabras	Apertium: permite prototipar y construir traductores automáticos eficientes sin necesidad de tener que comenzar su desarrollo desde cero.	-La automatización del proceso de traducción es particularmente compleja porque el programa se enfrenta a problemas como la ambigüedad. -Se necesita crear un nuevo sistema de traducción a partir de esta plataforma
Evaluación de un sistema de traducción automática basado en reglas o por qué BLEU sólo sirve para lo que sirve. [14]	Traducción del idioma español al euskera	Prototipo Matxin	-(HTER) (PostEdición) Un editor necesitaría cambiar 4 de cada 10 palabras para corregir la salida del sistema. -Traduce de español a uskera.
Reutilización de datos lingüísticos para la creación de un sistema de traducción automática para un nuevo par de lenguas. [2]	Traducción del idioma portugués al catalán	Una nueva herramienta de traducción de idioma basado en la plataforma Apertium	-Solo es traductor del idioma portugués al catalán En los diccionarios generados por apertium-dixtools hay dos tipos de errores de consistencia: -Los generados por el programa y los causados por los distintos criterios lingüísticos a la hora de desarrollar los diccionarios morfológicos de los traductores
Estudio comparativo de traductores automáticos en línea: Systran, Reverso y Google. [10]	Identificar el mejor sistema de traducción automática en línea	Google Systran Reverso	No se realiza una evaluación sobre textos paralelos para identificar la coherencia del texto de los traductores.

Tabla 3 Comparación de artículos y tesis relacionados.

Artículos y Tesis Relacionados	¿Cuál es el problema?	¿Qué herramientas se utilizaron?	¿Cuáles son las desventajas?
TESIS Sistema resumidor-traductor automático. [3]	Traducir textos en idioma inglés al español y resumirlos tomando en cuenta la calidad.	Herramienta de resumen: ENERTEX (FRESA) Herramienta de traductor: REVERSO (Sistema de calidad estadística)	1. El sistema está delimitado únicamente a documentos con carácter informativo, (noticias). 2. Solo traduce texto en inglés. 3. La herramienta ENERTEX, aunque fue evaluado con FRESA, no es la mejor herramienta de resumen en calidad. 4. Se observó un bajo desempeño del sistema para aquellos documentos que eran de pequeñas dimensiones (menores a 150 palabras de longitud). 5. Los resúmenes no son tan buenos cuando se realizan al 10% del documento original.
Columbia Newsblaster: Multilingual News Summarization on the Web. [15]	Este artículo aborda el problema de acceso de los usuarios a navegar por noticias de múltiples idiomas de varios sitios en el Internet.	Herramienta de resumen: COLUMBIA SUMMARIZER Herramienta de traductor: SYSTRAN	1. Los documentos que genera son en idioma inglés. 2. El artículo no habla de calidad del documento de salida. 3. El artículo es del año 2004

Capítulo 4
Metodología de Solución

Para la realización de este trabajo fue necesario diseñar una metodología, la cual consta de 7 fases divididas en 4 etapas como se muestra en la Figura 9, tomando en cuenta desde el análisis de los algoritmos de resumen y traducción automática hasta la implementación de los mismos.

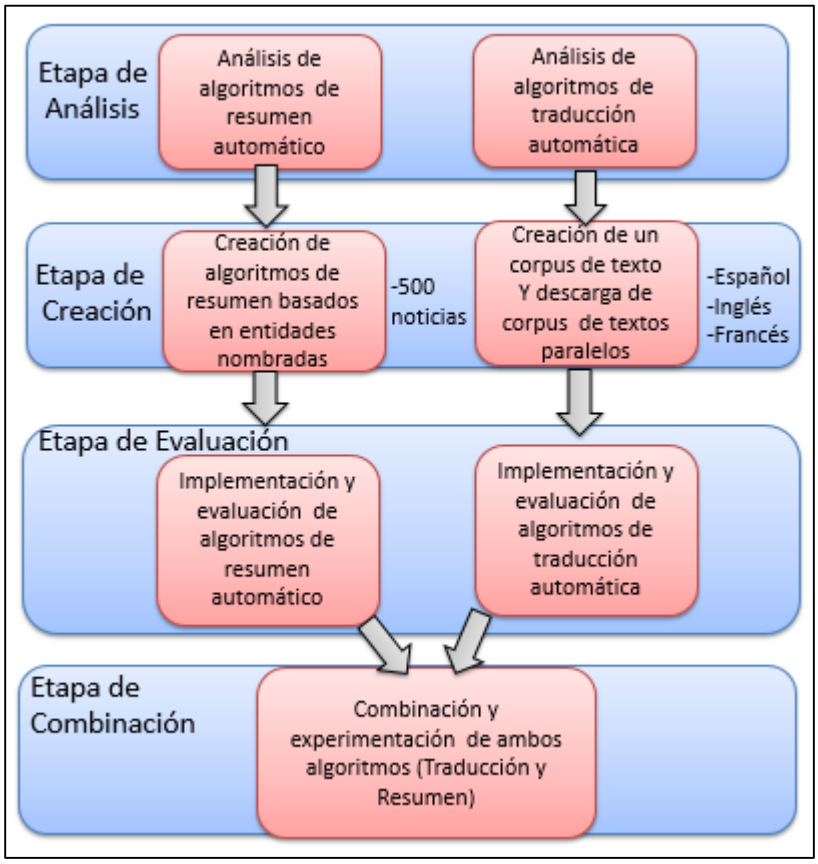


Figura 9 Metodología de solución

4.1 Etapa de análisis

Esta etapa engloba las fases de análisis, las cuales se describen a continuación.

4.1.1 Análisis de algoritmos de resumen automático

El análisis consiste en realizar una tabla comparativa (ver tabla 11) con base al estado del arte en donde se expongan las herramientas que utilizan los artículos sobre resumen automático y las desventajas que se enfrentan en sus resultados, por otro lado, se analizan las evaluaciones realizadas por los mismos para determinar que algoritmo de resumen automático de textos es el que genera mejores resultados y sobre que métrica se están evaluando, esta fase se completa con el capítulo Estado del Arte.

4.1.2 Análisis de algoritmos de traducción automática

El análisis de los traductores automáticos se realiza con base al estado del arte, el cual consiste en identificar el sistema de traducción automática que genere mejores resultados al momento de realizar una traducción de texto, este sistema de traducción también debe tener la capacidad de traducir los idiomas español, inglés y francés.

De acuerdo al capítulo Estado del Arte y con base a las evaluaciones que realiza el artículo “Estudio comparativo de traductores automáticos en línea: Systran, Reverso y Google” (ver punto 3.2.1) se determinó utilizar los sistemas de traducción automática Google, Systran y Reverso para evaluarlos mediante una métrica de evaluación Fresa que más adelante veremos en el capítulo 5 Implementación y diseño de esta tesis.

4.2 Etapa de creación

En esta etapa tenemos la creación de algoritmos de resumen automático de textos basados en entidades nombradas y la descarga de corpus para las evaluaciones.

4.2.1 Creación de un corpus de texto y descarga de corpus de textos paralelos

Para la conformación del corpus, fue necesario descargar alrededor de 500 noticias a través de diversos portales basados en ciencia y tecnología como “Huff post [16], eitb [17] y Noticias Google [18].

Para la extracción de noticias, se utilizó un módulo ocupando la biblioteca JSoup [8]. Esta analiza la estructura HTML de una página web para ubicar y extraer específicamente la información deseada.

La descarga del corpus se realizó para la evaluación de los algoritmos de resumen automático.

También se descargaron corpus de 50 líneas paralelos en idiomas español, inglés y francés [19] para la evaluación de los sistemas de traducción automática.

4.2.2 Creación de algoritmos de resumen basados en entidades nombradas

En esta fase se crearon 4 algoritmos de resumen automático de texto basados en entidades nombradas, se crearon algoritmos para la generación de líneas base y un extractor de información el cual consiste en concatenar múltiples documentos de texto plano en uno solo.

Los algoritmos que se crearon son los siguientes:

- NASON (New Automatic Summary based on prOper Nouns) Algoritmo basado en nombres propios
- NASVE (New Automatic Summary based on Verbs)

Algoritmo basado en verbos

- STONV (Summary Text based on prOper Nouns and Verbs)
  Algoritmo basado en nombres propios y verbos
- Algoritmo Híbrido
  EL algoritmo híbrido está basado en tres algoritmos (Nasve, Stonv y Artex)

4.3 Etapa de evaluación

En esta etapa se evaluaron los algoritmos de resumen automático y los sistemas de traducción automática.

4.3.1 Implementación y evaluación de algoritmos de resumen automático

En esta fase se evaluaron los algoritmos de resumen basados en entidades nombradas que se crearon junto con el mejor algoritmo de resumen automático Artex que se obtuvo con base al estado del arte para compararlos utilizando la herramienta de evaluación Fresa.

4.3.2 Implementación y evaluación de algoritmos de traducción automática

En esta fase se evaluaron los mejores sistemas de traducción automática basándose en el estado del arte, realizando pruebas mediante una métrica de evaluación para determinar qué sistema de traducción nos genera un mejor resultado en cuanto a sus traducciones en español, inglés y francés

4.4 Etapa de combinación

En esta última etapa se realizaron pruebas combinando el mejor algoritmo de resumen automático junto con el mejor sistema de traducción automática.

4.4.1 Combinación y experimentación de ambos algoritmos (Traducción y Resumen)

Para la combinación de un sistema de resumen y un sistema de traducción es necesario realizar pruebas para saber que nos conviene más al momento de procesar el texto, si primero realizar una traducción y después el resumen o primero realizar un resumen y después una traducción, para este caso se realizaron las dos pruebas y se evaluaron mediante una métrica de evaluación.

Capítulo 5
Diseño e Implementación

En este capítulo se describen las etapas de creación, evaluación y combinación de la metodología de solución.

5.1 Etapa de Creación

Para la etapa de creación se describe a continuación la fase “Creación de algoritmos de resumen basados en entidades nombradas”.

5.1.1 Instalación de Freeling

Se instaló la herramienta Freeling para ocupar sus datos de salida, esta herramienta realiza etiquetado POS, del cual nos interesa identificar nombres propios y verbos.
A continuación, se muestra un ejemplo del documento de salida que arroja Freeling.

```
1_01-11-2016 1_01-11-2016 Z 1
Errores errores NP00000 1
comunes común AQ0CP00 1
que que PR0CN00 0.550139
se se P00CN00 0.494509
cometen cometer VMIP3P0 1
en en SP 1
el el DA0MS0 1
' ' Fe 1
marketing marketing NCMS000 1
' ' Fe 1
móvil móvil AQ0CS00 0.765152
. . Fp 1
~ ~ Fz 1
2_01-11-2016 2_01-11-2016 Z 1
De de NP00000 1
coches coche NCMP000 1
autónomos autónomo AQ0MP00 0.661294
```

Figura 10 Ejemplo de los datos de salida que genera Freeling

5.1.2 Creación de un algoritmo para la extracción de información

El corpus que se creó está conformado por una gran cantidad de noticias de ciencia y tecnología, los cuales se encuentran en formato txt. Se creó un algoritmo de extracción de información, el cual tiene la función de extraer todas las noticias del corpus por líneas (estas líneas se definen por título, punto y signo de pregunta dentro del cuerpo de la noticia) y colocarlas en un solo documento txt, cada noticia tiene una estructura de identificadores para no perder el orden de su texto correspondiente.

Cada título tiene un número que lo identifica, el orden de identificador de las líneas del cuerpo de la noticia consiste en lo siguiente:

- Número de la línea,
- Número del identificador del título
- Nombre de la carpeta donde se encuentran las noticias (en este caso el nombre de la carpeta es la fecha de descarga del corpus)

Ejemplo del identificador de título:

1_01-11-2016 Titulo 1
2_01-11-2016 Titulo 2
3_01-11-2016 Titulo 3

Ejemplo del identificador de línea:

1_1_01-11-2016 primera línea que pertenece al título 1
2_1_01-11-2016 segunda línea que pertenece al título 1

3_2_01-11-2016 primera línea que pertenece al título 2
4_2_01-11-2016 segunda línea que pertenece al título 2

5_3_01-11-2016 primera línea que pertenece al título 3
6_3_01-11-2016 segunda línea que pertenece al título 3

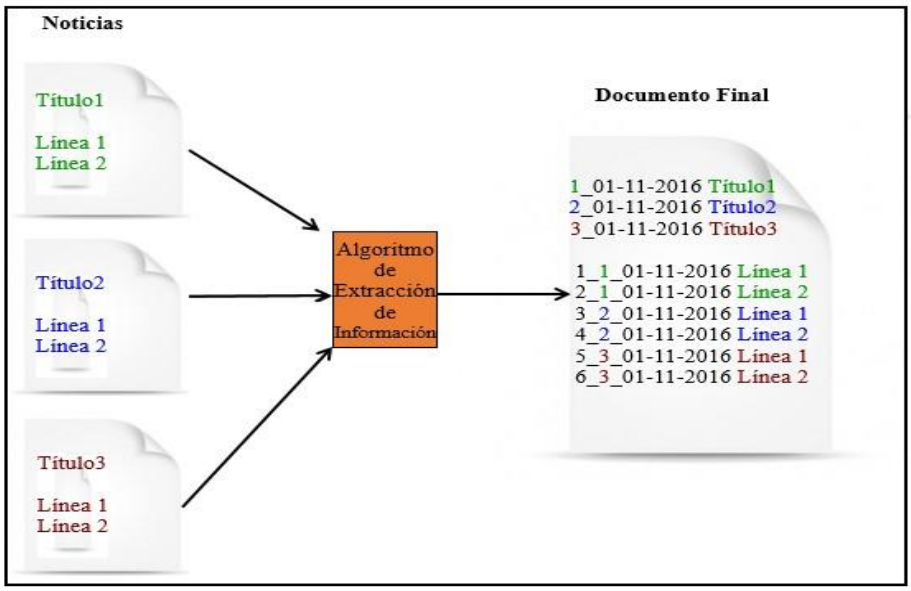


Figura 11 Estructura del algoritmo de extracción de información.

El objetivo de este algoritmo como se representa en la figura 11, es el de tener un corpus multi-documento, ya que al agrupar las noticias tenemos como entrada una gran cantidad de documentos el cuál, nos sirve para la entrada de cualquier algoritmo de resumen o traducción automática de textos.

5.1.3 Creación de algoritmo NASON (New Automatic Summary based on prOper Nouns) basado en nombres propios

Se creó un algoritmo de resumen automático de textos a partir de los datos etiquetados que devuelve Freeling. Con base al número de nombres propios que contiene cada línea, se fue asignando un peso a cada una de ellas, estos nombres propios se identificaron con la etiqueta (NP) que Freeling les asigna, el procedimiento del algoritmo de resumen es el siguiente:

1. EL algoritmo de resumen automático de texto Nason basado en nombres propios (NP) toma como entrada el archivo de la salida de Freeling, uniendo las palabras por línea y contando el número de NP que tiene cada una.

2. La cantidad de nombres propios que contiene cada frase es el peso que se le asignó a cada una, posteriormente se identificó la frase de mayor peso y se dividió el peso de cada frase entre el mayor obteniendo de esta forma el score normalizado.

Tabla 4 Ejemplo de score normalizado.

No. frase	***Palabras***	***de***	***las***	***frases***				***NP***	***NP/4***
1	**Peña**	es	el	presidente	de	**México**		2	0.5
2	**Carmen**	y	**Felipe**	se	casarán	en	París	3	0.75
3	**Luis**	**Juan**	**Roberto**	y	**Karen**	viajarán		**4**	1

En el ejemplo anterior se puede observar que el número más alto es el 4 que corresponde a la tercera frase, en este ejemplo por medio del número 4 se obtiene el score normalizado de cada frase.
Se genera un corpus que contiene cada frase con su score normalizado, pero sin ponderar, esto, para que el corpus no pierda el orden original.

3. Se realiza la ponderación de frases, creando un nuevo documento del corpus, pero ordenando las frases de mayor a menor de acuerdo al peso que tiene cada una, generando un documento ponderado de todo el corpus.

4. Ahora se obtiene el 30% del corpus ponderado en un nuevo documento obteniendo las primeras líneas, formando el resumen con las líneas de mayor ponderación.
5. El siguiente paso consiste en la ordenación de las frases de acuerdo al número de identificador que tiene cada una de ellas y así obtener un resumen coherente.
6. Por último, se elimina el identificador de cada frase para tener un texto limpio.

En total se obtienen como salida los siguientes documentos:

-Corpus de las frases con su score normalizado.

-Corpus ponderado de las frases.

-Resumen del 30% del corpus

5.1.4 Creación de algoritmo de resumen automático de texto Nasve y Stonv

Al igual que el algoritmo de resumen basado en nombres propios(Nason) se creó un algoritmo de resumen basado en verbos (Nasve) y uno basado en nombres propios más verbos (Stonv) siguiendo el mismo procedimiento.

Estos algoritmos de resumen automático obtienen los mismos documentos que el anterior:

-Corpus de las frases con su score normalizado.

-Corpus ponderado de las frases.

-Resumen del 30% del corpus

5.1.5 Creación de algoritmo de resumen automático de textos híbrido

Se creó un algoritmo hibrido con base a los 3 mejores algoritmos de resumen de acuerdo a la evaluación de FRESA, estos algoritmos son:

-Artex

-Stonv

-Nasve

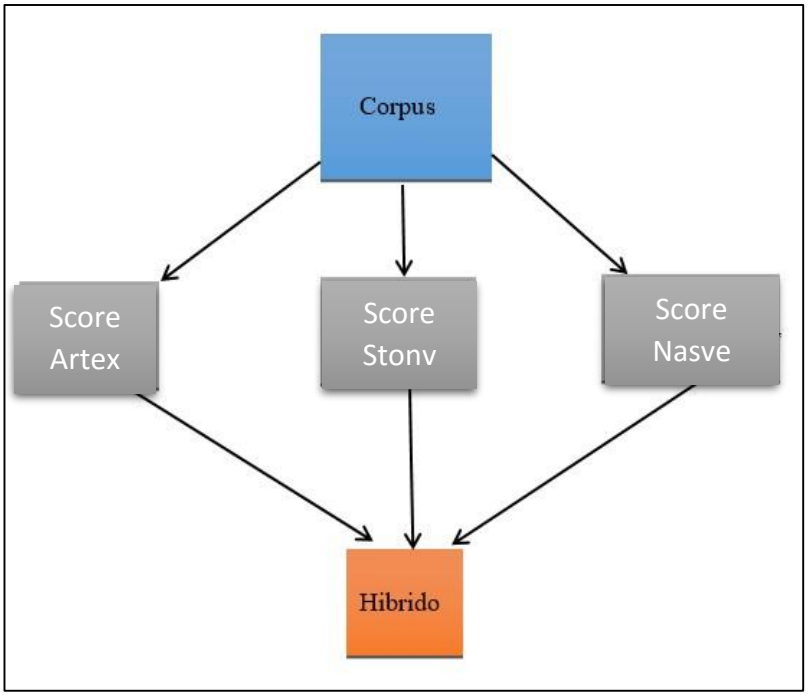


Figura 12 Algoritmo híbrido

El algoritmo híbrido utiliza el score de cada uno para generar su propio score como se muestra en la figura 12.

Tabla 5 Ejemplo de la obtención del score híbrido.

No. de frase	*Score Artex*	*Score Nasve*	*Score Stonv*	*Score Híbrido*
1	0.89	0.85	0.90	0.79
2	0.76	0.80	0.86	0.80
3	0.68	0.70	0.72	0.7

(ScoreArtex + ScoreNasve + ScoreStonv) /3 = score híbrido

5.1.6 Algoritmos para la creación de baseline

Un baseline es un resumen del corpus fuente tomando líneas aleatorias, pero del mismo tamaño del generado por el algoritmo de resumen automático de textos a evaluar, en este caso Artex

Artex cuenta con un segmentador, el cual segmenta por líneas el corpus generado por el extractor de información, a partir de este corpus segmentado se creó un algoritmo para la generación del baseline, el cual cuenta el número de

líneas totales que tiene el corpus y mediante una regla de tres se obtiene el número de líneas que corresponde al 30% del corpus que es el porcentaje de resumen que se ocupa como base en Artex.

Figura 13 Generación de baseline con líneas aleatorias.

Como se puede observar en la figura 13 nos indica el número total de líneas que contiene el corpus y el número de líneas que corresponden al 30% para generar un resumen con líneas aleatorias. Se crearon 3 distintas baseline aleatorias para las pruebas.

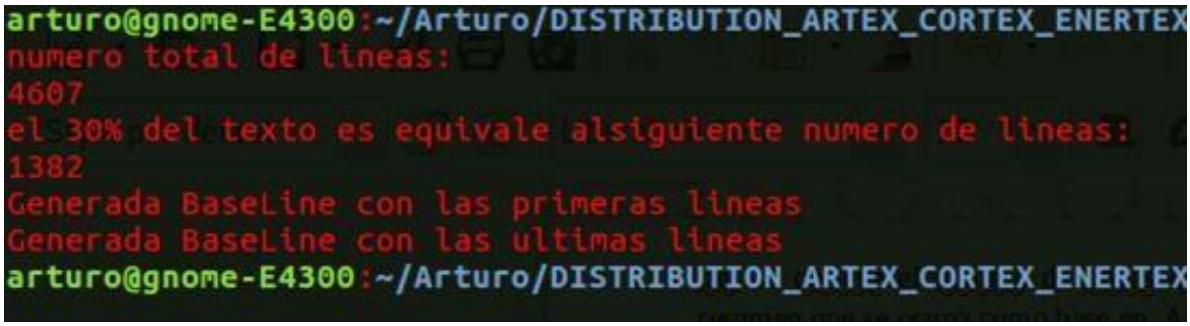


Figura 14 Generación de baseline con las primeras y últimas líneas del corpus.

Con fines de evaluación también se creó un algoritmo que genera dos tipos de baseline diferentes a los aleatorios, el primer baseline se forma a partir de las primeras líneas del corpus y el segundo con las últimas líneas como se muestra en la figura 14, teniendo un total de 5 baseline, 3 aleatorios y los dos que se mencionan aquí.

5.2 Etapa de Evaluación

En esta etapa se presenta la herramienta de evaluación que se utilizó, el mejor algoritmo de resumen automático de texto que se obtuvo con base al estado del arte y las evaluaciones realizadas a los algoritmos de resumen automático de texto y a los sistemas de traducción automática.

5.2.1 Herramienta de evaluación Fresa

Para evaluar la calidad de los resúmenes generados por los algoritmos de resumen automático de texto se utilizó la herramienta Fresa (*Framework for Evaluating Summaries Automatically*), esta herramienta fue desarrollada en el Laboratorio de Informática de Aviñón (LIA) de la Universidad de Aviñón en el año 2010 por el Doctor Juan Manuel Torres Moreno. [9]

Fresa calcula divergencias entre los resúmenes y el origen del documento, ya que integra un conjunto de estadísticas que permiten realizar una comparación de distribuciones de probabilidad.

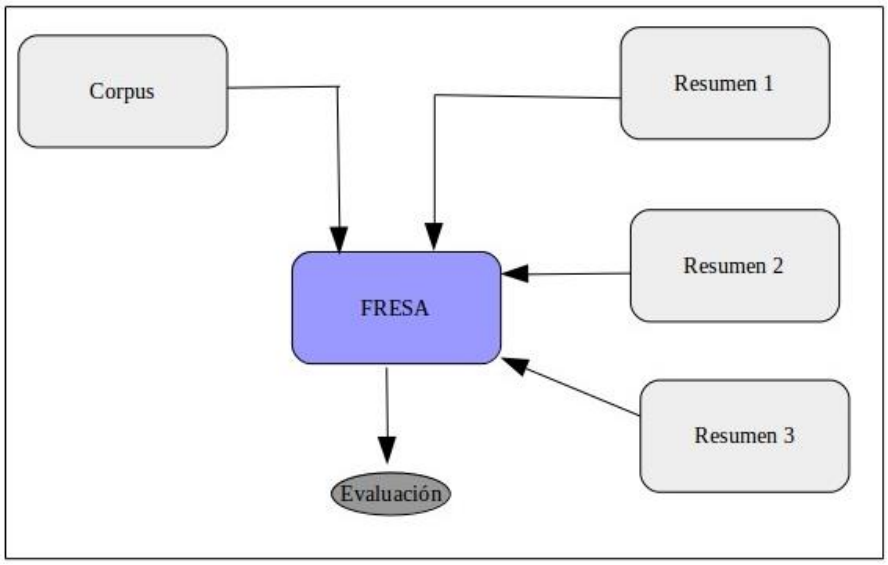


Figura 15 Representación de las entradas y la salida de FRESA.

En la figura 15 se representa la estructura de Fresa respecto a las entradas y su salida que es la evaluación.

Este sistema evalúa los resúmenes mediante N-gramas, -Bi-gramas y Bi-gramas con huecos, también calcula la media de los 3.

Se eligió este sistema de evaluación dado que el desempeño de FRESA ha sido comparado con otras herramientas como Rouge [20], sin embargo, FRESA no ocupa referencia humana para la evaluación de calidad de los resúmenes, esto beneficia al momento de realizar evaluaciones ya que solo se depende de una computadora. Cabe mencionar que Fresa trabaja en sistema operativo Linux.

5.2.2 Instalación de Artex

Se utilizó el algoritmo de resumen Artex como referencia para la evaluación y comparación de los algoritmos Nasve, Nason, Stonv e híbrido ya que ha sido evaluado junto con otros algoritmos de resumen automáticos de textos como Cortex y Enertex) por el algoritmo FRESA (sin referencias de humanos) [1].

5.2.3 Evaluación de resumen de texto

A continuación, se presentan los resultados sobre la evaluación de los algoritmos de resumen automático de texto.

Tabla 6 Evaluación Fresa (Resumen).

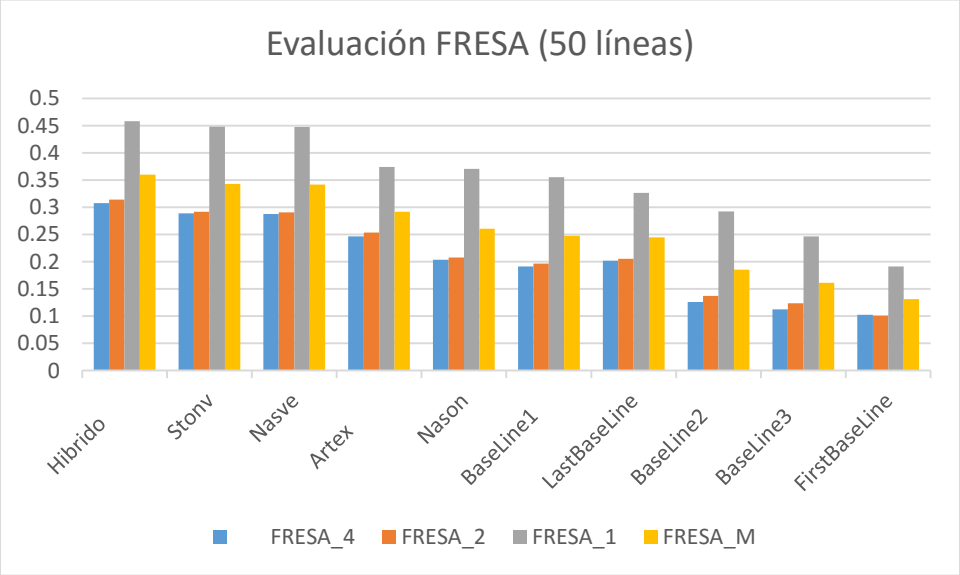
Sistema	*Bi-gramas con huecos (4)*	*Bi-gramas (2)*	*N-gramas (1)*	*Media (M)*
Stonv	**0.30518**	**0.31801**	**0.60303**	**0.40874**
Híbrido	0.30456	0.31679	0.58802	0.40313
Nasve	0.29516	0.30850	0.57619	0.39328
Artex	0.30025	0.30532	0.54338	0.38298
Nason	0.21644	0.22362	0.55119	0.33042
LastBaseLine	0.18688	0.18207	0.43943	0.26946
FirstBaseLine	0.17778	0.17095	0.42721	0.25865
BaseLine3	0.15398	0.16448	0.45461	0.25769
BaseLine2	0.15001	0.16003	0.45488	0.25498
BaseLine1	0.15025	0.16070	0.44552	0.25216

Como hemos podido observar en la tabla de resultados, se evaluaron los 4 algoritmos de resumen automático de texto con un corpus de 500 noticias y los cuales generan resúmenes de calidad de acuerdo a la evaluación Fresa [9]. Posicionándose en primer lugar el algoritmo Stonv nos demuestra los resultados más altos en las 4 categorías, (Bi-gramas con huecos (4), Bi-gramas (2), N-gramas (1) y la Media(M)) seguido se encuentra el algoritmo Híbrido con una ligera diferencia en sus resultados con respecto a Stonv. Observemos que Artex Ocupa el 4to lugar en la tabla de evaluación siendo uno de los mejores algoritmos de acuerdo

al Estado del Arte, demostrando que 3 de los 4 algoritmos creados entre ellos Stonv, Híbrido y Nasve) generan buenos resultados. De acuerdo a esta evaluación se ha elegido el algoritmo Stonv para implementarlo con la traducción automática.

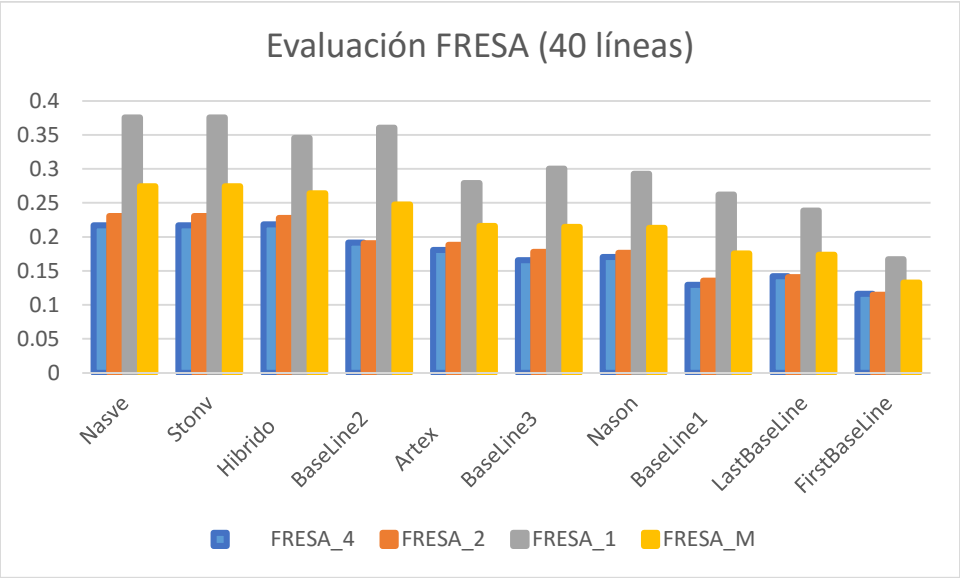
5.2.4 Evaluación de algoritmos de resumen con textos cortos

Se realizaron nuevas evaluaciones a los algoritmos de resumen automático para observar cómo se comporta el algoritmo Stonv con respecto a textos cortos, los textos con los que se realizaron las pruebas están conformados por 50, 40 y 30 líneas cada uno y los resultados se muestran a continuación:



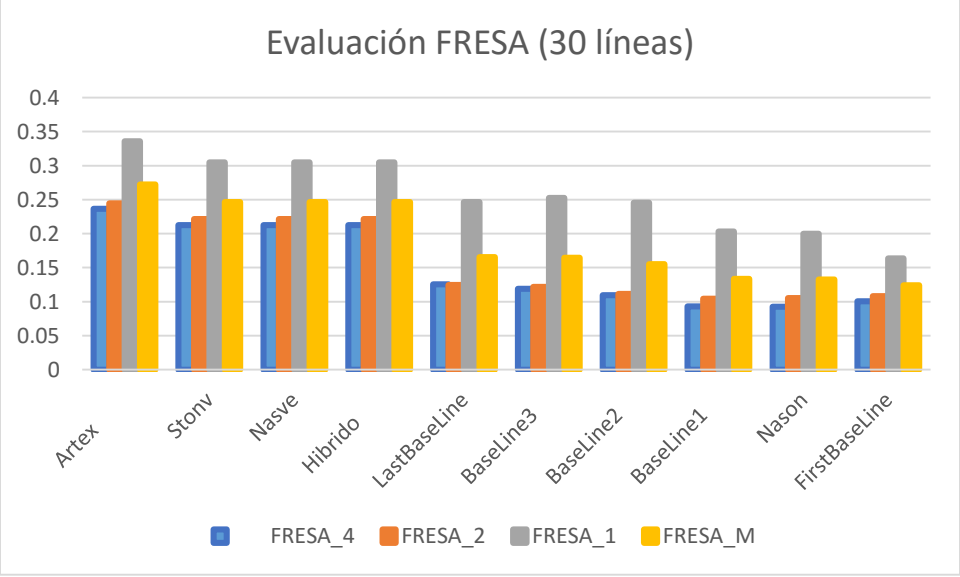
Gráfica 1 Evaluación de resumen con 50 líneas

En la gráfica 1 podemos observar que los resultados con un texto corto de 50 líneas, favorecen al algoritmo híbrido con una media de 0,36027 mientras que el algoritmo Stonv se posiciona en tercer lugar con una media de 0,34219. Para ver los resultados con mayor detalle ver anexo1.



Gráfica 2 Evaluación de resumen con 40 líneas

En la evaluación con el texto de 40 líneas el algoritmo basado en verbos (Nasve) se posiciona en primer lugar mientras que el algoritmo Stonv sube a la segunda posición con una diferencia mínima con respecto a Nasve.



Gráfica 3 Evaluación de resumen con 30 líneas

En la evaluación con 30 líneas vemos que Artex obtiene la primera posición, sin embargo, Stonv se mantiene en la segunda posición de la gráfica.

5.2.5 Conclusión de las evaluaciones de resumen de textos cortos

Como podemos observar en las evaluaciones anteriores, el comportamiento del algoritmo Stonv en textos cortos no fue el más optimo, sin embargo, se posiciona entre los mejores resultados de las evaluaciones, por otro lado, cabe mencionar que para esta tesis se eligió el algoritmo Stonv ya que obtiene mejores resultados con respecto a la evaluación de textos grandes como el corpus conformado por 500 noticias.

5.3 Evaluación de traducción de texto

Para la evaluación de la calidad de los sistemas de traducción automática en línea se utilizó la herramienta Fresa (*Framework for Evaluating Summaries Automatically*) [9] y los corpus paralelos de 50 líneas en idiomas español, inglés y francés descargados anteriormente. El procedimiento de evaluación se realizó de la siguiente manera:

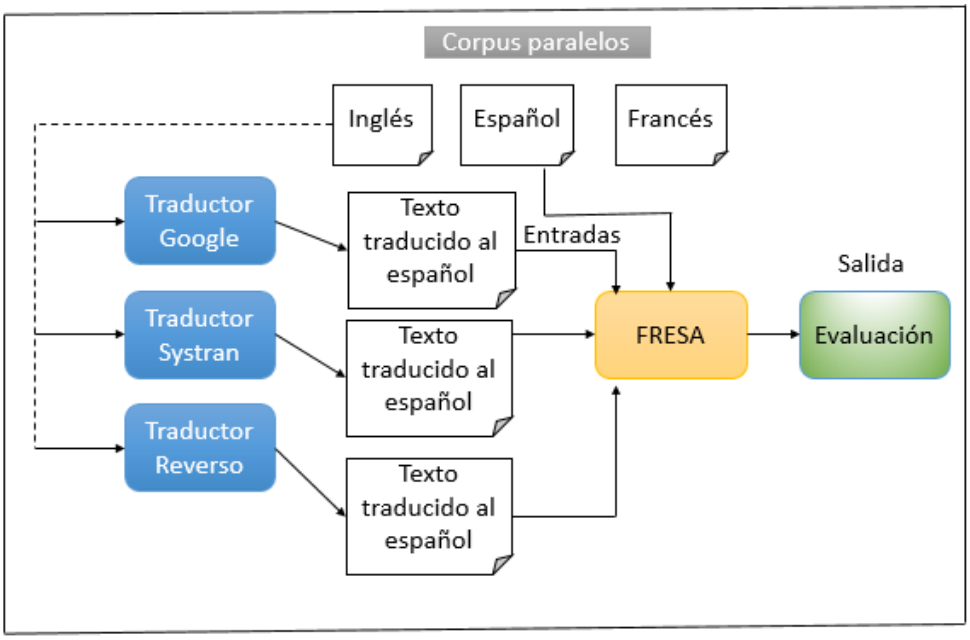


Figura 16 Proceso de evaluación de traductores con corpus paralelos.

El procedimiento que se muestra en la figura 16, consistió en introducir el corpus paralelo en inglés a los traductores Google, Systran y Reverso, teniendo como salida 3 documentos con texto traducido al idioma español, los cuales fueron la entrada al sistema de evaluación Fresa junto con el corpus paralelo en español, el cual jugó el papel de documento fuente.

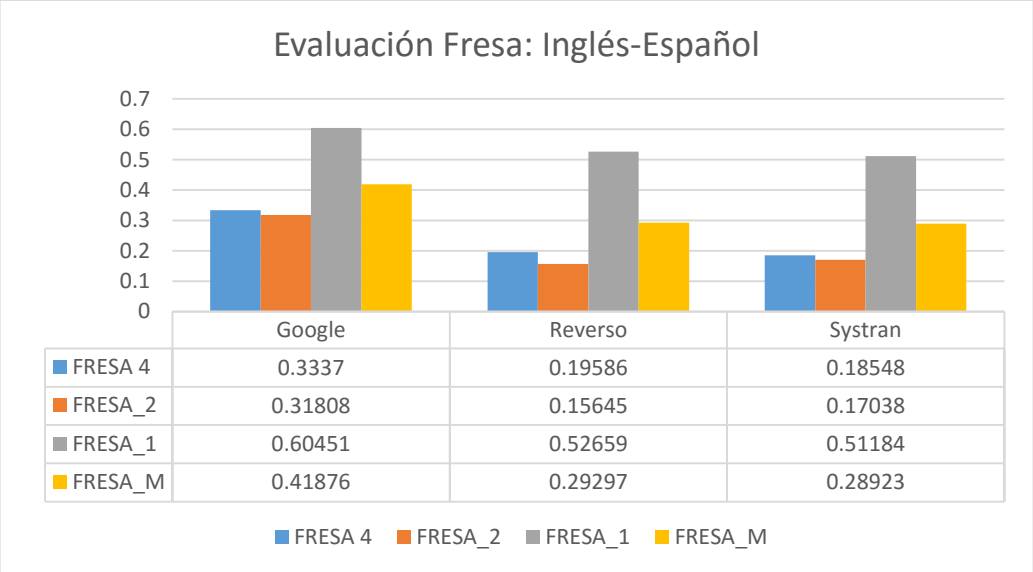
Fresa nos entrega como salida la evaluación de las traducciones con respecto al corpus paralelo en español, este mismo proceso se realizó con los 3 idiomas traduciendo los corpus de la siguiente manera:

- Traducción del español al inglés y del inglés al español
- Traducción del francés al inglés y del inglés al francés
- Traducción del español al francés y del francés al español

De esta forma se evaluó la traducción realizada con los corpus paralelos en el idioma correspondiente.

Recordemos que Fresa evalúa mediante N-gramas (FRESA_1), bi-gramas (FRESA_2), bi-gramas con huecos (FRESA_4) y obtiene la media de estos 3 (FRESA_M) y está basado en las divergencias *Kullback-Leibler* (KL)[21] y *Jensen-Shanon* (JS)[21].

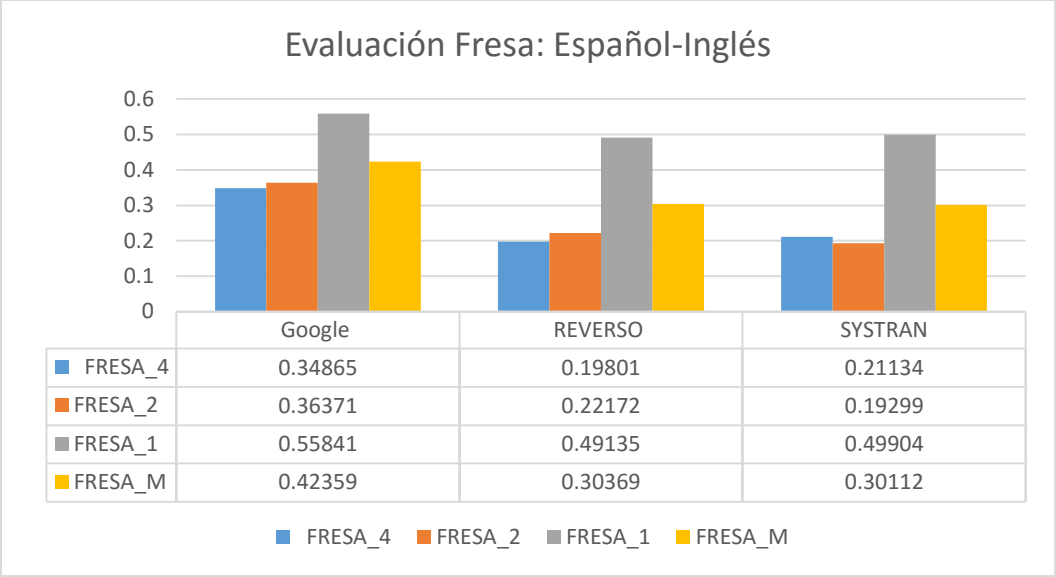
A continuación, se presentan los resultados obtenidos de las 6 evaluaciones realizadas:



	Google	Reverso	Systran
FRESA 4	0.3337	0.19586	0.18548
FRESA_2	0.31808	0.15645	0.17038
FRESA_1	0.60451	0.52659	0.51184
FRESA_M	0.41876	0.29297	0.28923

Gráfica 4 Evaluación de traducción inglés a español.

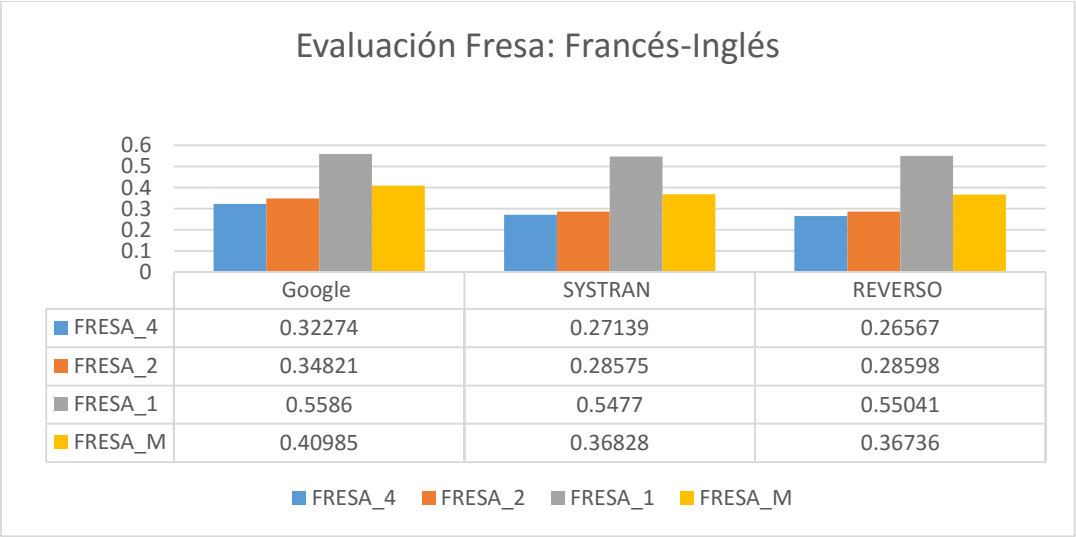
En la gráfica 4 podemos observar que Google obtiene mejores resultados en cuanto a la traducción de inglés al español.



	Google	REVERSO	SYSTRAN
FRESA_4	0.34865	0.19801	0.21134
FRESA_2	0.36371	0.22172	0.19299
FRESA_1	0.55841	0.49135	0.49904
FRESA_M	0.42359	0.30369	0.30112

Gráfica 5 Evaluación de traducción español a inglés.

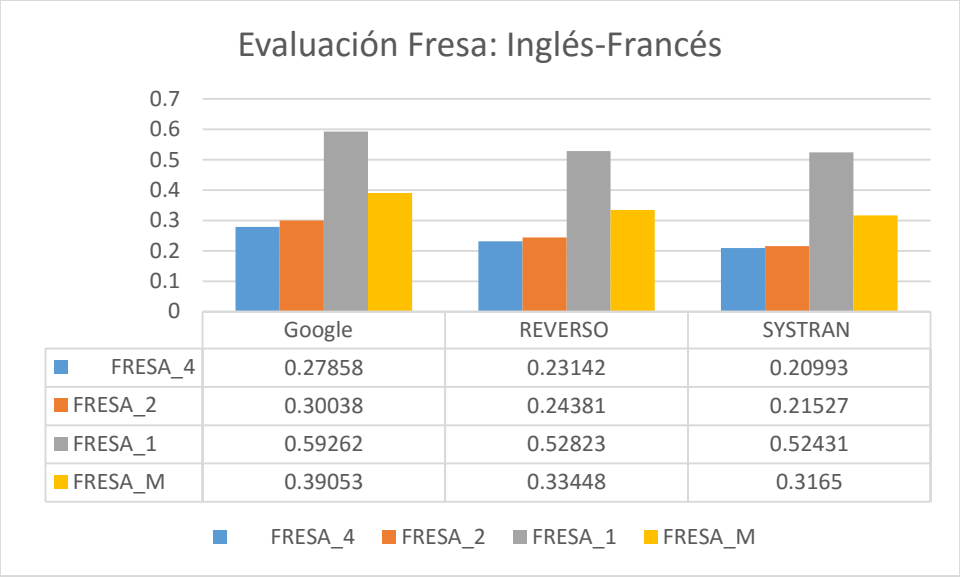
En la gráfica 5 tenemos la traducción a la inversa, del español al inglés y podemos observar que el traductor Google sigue teniendo el primer lugar en cuanto a los resultados.



	Google	SYSTRAN	REVERSO
FRESA_4	0.32274	0.27139	0.26567
FRESA_2	0.34821	0.28575	0.28598
FRESA_1	0.5586	0.5477	0.55041
FRESA_M	0.40985	0.36828	0.36736

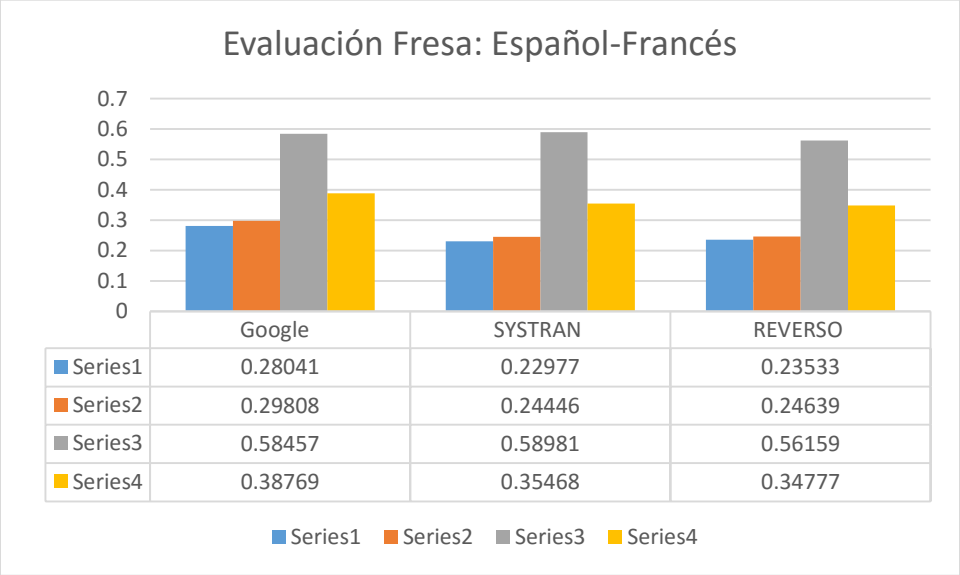
Gráfica 6 Evaluación de traducción francés a inglés.

En la gráfica 6 se puede observar que los resultados se emparejan un poco dado que ahora la traducción se está realizando del francés al inglés, sin embargo, el traductor Google sigue manteniéndose con los mejores resultados en la evaluación.



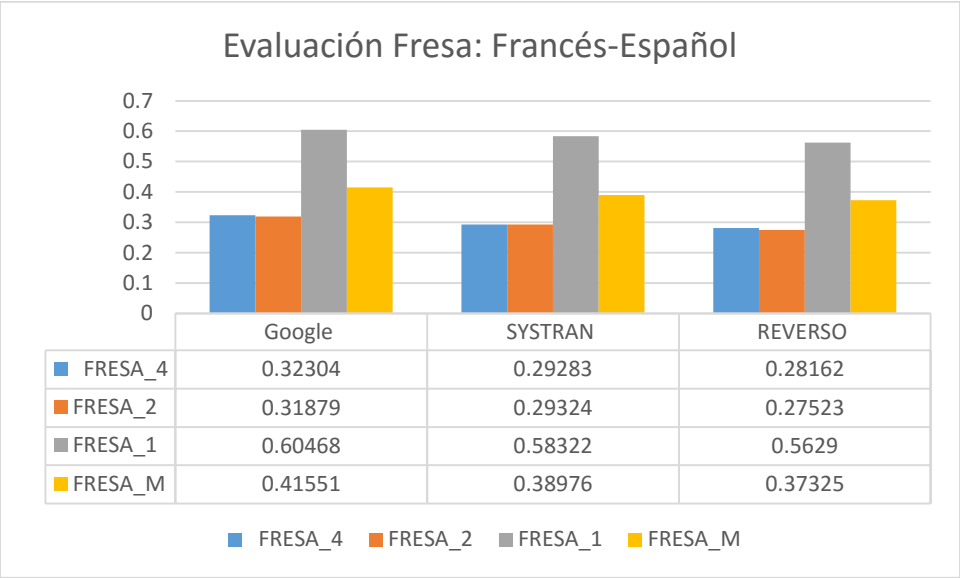
Gráfica 7 Evaluación de traducción inglés a francés.

Para la evaluación de inglés a francés el traductor Google lleva ligeramente un mejor resultado a comparación del traductor Reverso, sin embargo, el traductor Systran ha ocupado el último lugar a excepción de la evaluación en la gráfica 6.



Gráfica 8 Evaluación de traducción español a francés.

En esta evaluación se puede observar como Systran iguala a Google en la evaluación por n-gramas, pero en la media Google obtiene una ligera ventaja. Recordemos que la media es el promedio de los n-gramas, bi-gramas y bi-gramas con huecos.



Gráfica 9 Evaluación de traducción francés a español.

En la traducción del francés al español los resultados muestran una ligera diferencia entre los traductores, sin embargo, Google nos sigue dando mejores resultados en los cuatro métodos de evaluación que utiliza Fresa.

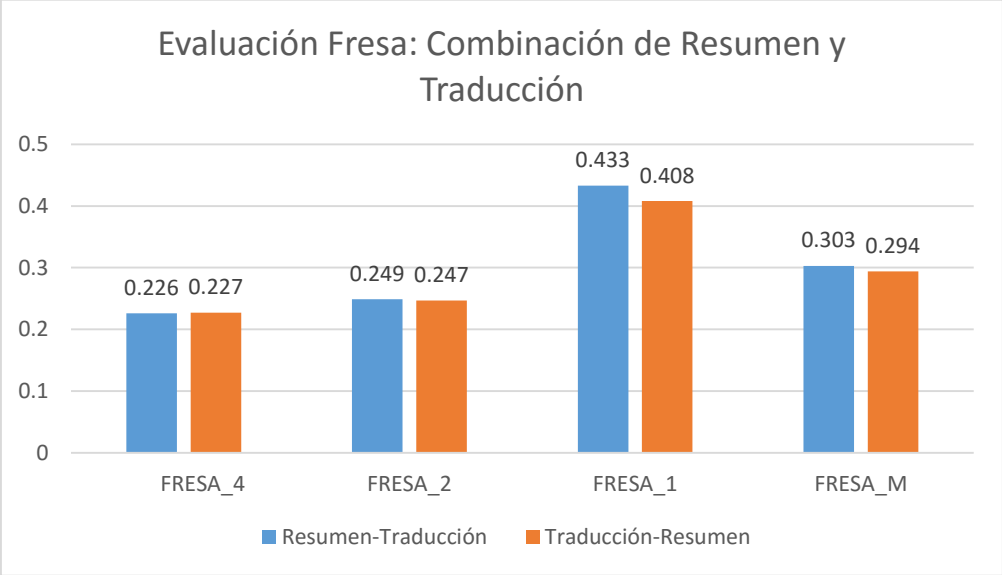
5.4 2.5 Conclusión de la evaluación de los sistemas de traducción

Como se ha demostrado en las evaluaciones anteriores con los 3 idiomas francés, español e inglés, Google es el traductor que genera mejores resultados a comparación de Systran y Reverso, por lo tanto, se ha decidido utilizar el traductor automático Google para combinarlo con el algoritmo de resumen automático de texto Stonv (Summary Text base on prOper Nouns and Verbs).

5.5 Etapa de combinación

Para realizar la combinación de ambos algoritmos, resumen y traducción, se realizó una evaluación para determinar la mejor forma de combinarlos, es decir, se comprobó qué combinación da mejores resultados, si primero realizar el resumen y después traducirlo, o primero traducir el texto y después realizar el resumen.

A continuación, se presenta una gráfica evaluando los resultados de las combinaciones de los sistemas de resumen y traducción mencionadas anteriormente.



Gráfica 10 Evaluación de los sistemas de resumen y traducción automática combinados.

En la gráfica 10 se demostró que obtenemos mejores resultados realizando primero el resumen y por consiguiente la traducción automática.
NOTA: Recordemos que el algoritmo Stonv realiza resumen de tipo extractivo, lo que significa que el resumen es generado con las frases originales, es decir, no modifica el texto al momento de generarlo, por lo tanto, aplicar el resumen primero nos dará mejores resultados ya que la traducción automática solo traducirá el texto resumido y no el corpus fuente. Por otro lado, si la traducción automática se realizara primero, esta modificaría el texto fuente y es más probable que se generen errores al momento de procesarlo.

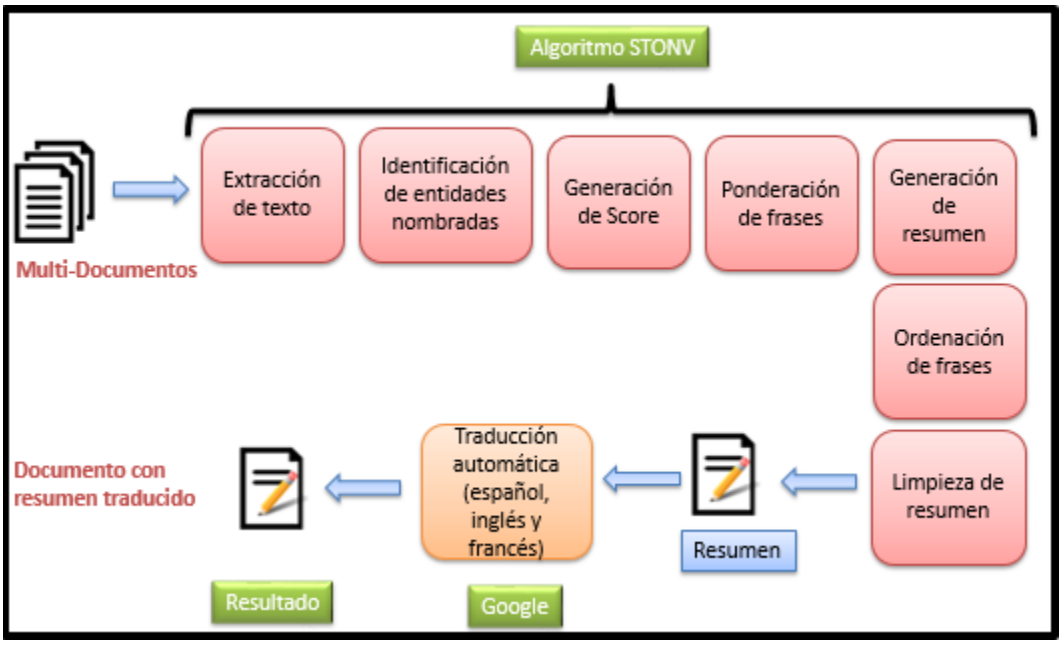


Figura 17 Implementación de resumen y traducción.

En la figura 17 observamos la estructura que se siguió para la implementación de resumen y traducción.
A continuación, se describen los pasos de la figura 17:

- Tenemos como entrada un corpus conformado por varios documentos, estos documentos pueden estar en inglés, francés, español o en idiomas combinados.
- Se realiza la extracción de frases por punto, signo de pregunta y por títulos.
- Se segmenta el texto por entidades nombradas (nombres propios y verbos) que son las entidades que utiliza el algoritmo Stonv.
- Con base al número de entidades por frase genera un score normalizado
- Realiza una ponderación de frases por su score normalizado.
- Realiza el resumen.
- Ordena las frases por su identificador único (esto para no perder coherencia en el resumen al momento de leerlo).
- Se eliminan los identificadores de las frases para que el texto quede limpio.
- El resumen se ingresa al traductor Google para traducirlo en uno de los 3 idiomas (inglés, francés o español).
- Se obtiene resumen traducido.

Capítulo 6
Pruebas y Resultados

En este capítulo, se presentan las pruebas con el objetivo de comprobar y validar la metodología, así como también analizar los resultados obtenidos en la combinación de resumen y traducción.

6.1 Evaluación de resumen automático Stonv

Se realizó la evaluación al algoritmo de resumen automático de texto Stonv (***S**ummary **T**ext based on pr**O**per **N**ouns and **V**erbs*) el cual identifica nombres propios y verbos, para comprobar la eficiencia del mismo mediante las medidas de *precisión, recall y f-measure*.

Se contaron de manera manual y por separado los nombres propios y verbos de cada frase, utilizando 100 de ellas para las pruebas (ver anexo 3). Los verbos identificados se analizaron con el diccionario RAE (Real Academia Española) [22] y mediante un experto para comprobar que los verbos identificados se usen como tal en el contexto de la frase, de lo contrario se toma como un falso positivo.

Enseguida se muestra las fórmulas y los resultados de las evaluaciones:

- Precisión = TP/ (TP+FP)
- Recall = TP/ (TP+FN)
- F-measure = (2*precision * recall) / (precision + recall)

Tabla 7 Medidas de evaluación a nombres propios.

Nombres propios		Valor real		100 frases	
		False	True		
Valor	Negative	FN 0	TN 5	Precisión	0,909090909
Stonv	Positive	FP 5	TP 50	Recall	0,909090909
				F-measure	0,909090909

En la tabla 17 se muestra que los resultados se mantienen en un 90% demostrando que la identificación de nombres propios es buena, en cuanto a el total de nombres propios reales que se identificaron en las 100 frases fueron 55, mientras que los predichos por el algoritmo fueron 50.

Tabla 8 Medidas de evaluación a verbos.

Verbos		Valor real			
		False	True	100 frases	
Valor	Negative	FN 0	TN 18	Precisión	0,939306358
Stonv	Positive	FP 21	TP 325	Recall	0,947521866
				F-measure	0,943396226

En la tabla 18 se muestra un resultado del 94% en F-measure que es el resultado de la precisión y el recall demostrando que tiene un mejor resultado la identificación de verbos, el número total de verbos reales fue de 343 y los predichos por el algoritmo fueron 325 en las 100 frases, como podemos ver son más los verbos que los nombres propios.

6.2 Generación de resumen por el algoritmo Stonv

En esta sección se realizó un seguimiento manual sobre el procedimiento del algoritmo Stonv para la generación de resumen mediante el score y la ponderación de las frases como se muestra a continuación.

```
0.222222222222   1_2017 ¿ están preparadas nuestras instituciones básicas ante un ciberataque ?.
0.666666666667   1_1_2017 el pasado_viernes_21_de_octubre distintas empresas de tanto relieve como Spotify, NYT o Twitter sufrieron un ataque
que las dejó paralizadas.
0.0   2_1_2017 casi.
0.333333333333   3_1_2017 las explicaciones que ofrecieron de el asunto distintos medios de comunicación y, sobre todo, distintos expertos
eran para poner nos los pelos de punta.
0.333333333333   4_1_2017 incluso aunque no las entendiéramos - o tal_vez por ello - estaban lejos_de tranquilizar nos.
0.444444444444   5_1_2017 parecía que la Tercera_Guerra_Mundial había empezado ; pero a_través_de nuestros cables sujetos a la red y nuestros
aparatos más próximos : el ordenador, la impresora y cosas así.
0.555555555556   6_1_2017 se hablaba de que sólo había sido una especie de prueba, de entrenamiento, y que el próximo podría ir contra
instituciones como hospitales o bancos.
0.111111111111   7_1_2017 a_partir_de ahí, la imaginación se dirigía hacia el apocalipsis.
0.222222222222   8_1_2017 hasta ahora, la mayor parte de el imaginario apocalíptico de el cine o la literatura ponía como catástrofe
originaria un desgraciado accidente nuclear, un ataque extraterrestre o una gran avería climática.
0.888888888889   9_1_2017 pocas veces era el producto de un ataque cibernético en el que los resultantes de el mismo habían sido producto de
unos dígitales que, armados con afán destructivo infinito, habían dormido en nuestros electrodomésticos hasta que alguien - no se sabe quién -
los despertó.
0.555555555556   10_1_2017 sí será el apocalipsis si logran alcanzar instituciones como, por_ejemplo, la Seguridad_Social o Hacienda.
0.555555555556   11_1_2017 podrían desaparecer los datos y nadie cobraría su pensión, porque habría desaparecido de la documentación.
0.444444444444   12_1_2017 cada uno tendría que demostrar todo lo que había cotizado durante todos sus años de vida laboral.
0.222222222222   13_1_2017 ¡ Casi nada ! Desde_luego que se solucionaba el problema de el acelerado adelgazamiento de la caja de las
pensiones ; pero de la peor forma posible.
0.666666666667   14_1_2017 lo mismo podría pasar con Hacienda, aun cuando aquí, bastantes podrían salir beneficiados.
0.222222222222   15_1_2017 eso hablando de instituciones públicas que, a el menos, parecen cibernéticamente seguras.
0.111111111111   16_1_2017 a el menos, eso espero.
0.111111111111   17_1_2017 pero pensemos en otras, como nuestras universidades.
1.0   18_1_2017 supongamos que un ciberataque borra toda huella de los archivos universitarios ; ¿ cómo se podrá mostrar quien tiene el título
para ejercer una profesión, si antes no ha tenido la ventura de conseguir lo en el viejo papel ?.
0.555555555556   19_1_2017 ¿ Tendrá que ir, profesor por profesor, solicitando que le recordara su paso por las aulas hace ya cinco, diez o
quince años ?.
0.333333333333   20_1_2017 todo lo anterior tal_vez no sea suficiente para generar esos lúgubres escenarios catastrofistas ; pero piensen en
un ciberataque a el sistema financiero, a cada uno o la mayor parte de los grandes bancos.
0.111111111111   21_1_2017 un borrado de nuestros ahorros o de nuestras inversiones.
```

Figura 18 Generación de score por frase.

En la figura 18 se muestra el score generado por el algoritmo de resumen a cada una de las frases de acuerdo al número de nombres propios y verbos que contienen. Las frases marcadas con una línea color azul, muestran las frases con un score más alto, en este ejemplo nos basaremos en la línea con su identificador 18_1_2017 (línea color naranja) la cual tiene el score más alto de 1.0. Esta línea es utilizada para formar parte del resumen automático por su alto score.

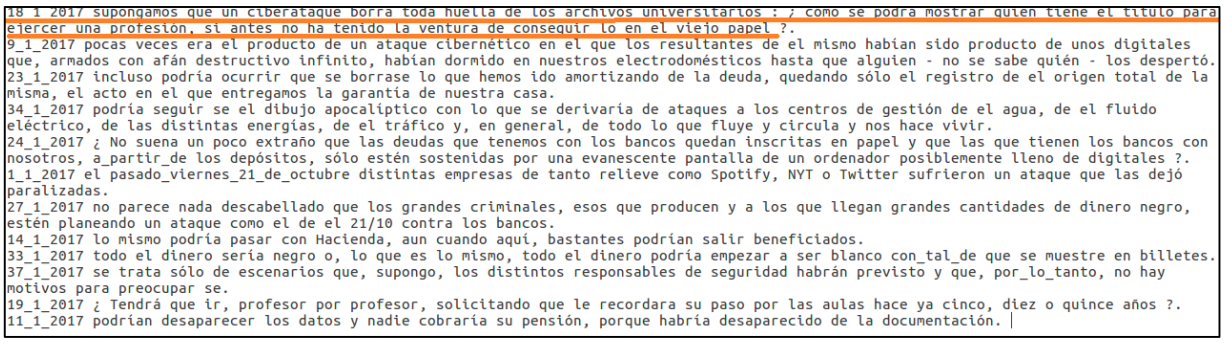


Figura 19 Ponderación de frases.

En la figura 19 tenemos las frases ponderadas por score, como podemos observar la rase que usamos como ejemplo se posiciona arriba ya que es la que tiene el score más alto.

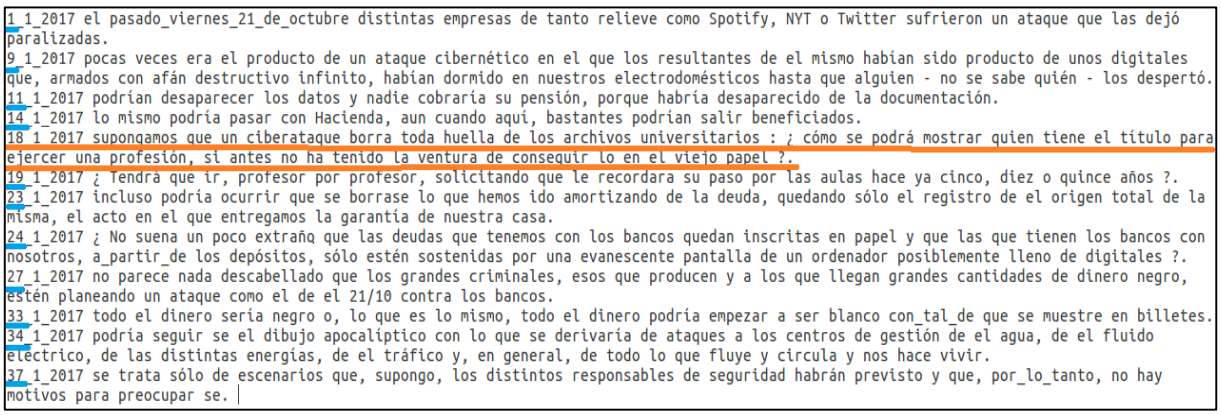


Figura 20 Resumen generado y ordenado.

En la figura 20 tenemos el resumen generado y se puede observar que la frase marcada de color naranja, forma parte del resumen final, sin embargo, aunque esta frase tuvo el score más alto no se encuentra en la primera línea, ya que las frases se ordenan mediante su número identificador (marcado de color azul) para generar un resumen coherente con base al orden de cada frase.

```
¿Están preparadas nuestras instituciones básicas ante un ciberataque?

El pasado viernes 21 de octubre distintas empresas de tanto relieve como Spotify, NYT o Twitter sufrieron un ataque que las dejó paralizadas.
Pocas veces era el producto de un ataque cibernético en el que los resultantes de el mismo habían sido producto de unos digitales que, armados
con afán destructivo infinito, habían dormido en nuestros electrodomésticos hasta que alguien - no se sabe quién - los despertó.
Podrían desaparecer los datos y nadie cobraría su pensión, porque habría desaparecido de la documentación.
Lo mismo podría pasar con Hacienda, aun cuando aquí, bastantes podrían salir beneficiados.
Supongamos que un ciberataque borra toda huella de los archivos universitarios : ¿ cómo se podrá mostrar quien tiene el título para ejercer
una profesión, si antes no ha tenido la ventura de conseguir lo en el viejo papel ?.
¿ Tendrá que ir, profesor por profesor, solicitando que le recordara su paso por las aulas hace ya cinco, diez o quince años ?.
Incluso podría ocurrir que se borrase lo que hemos ido amortizando de la deuda, quedando sólo el registro de el origen total de la misma, el
acto en el que entregamos la garantía de nuestra casa.
¿ No suena un poco extraño que las deudas que tenemos con los bancos quedan inscritas en papel y que las que tienen los bancos con nosotros, a
partir de los depósitos, sólo estén sostenidas por una evanescente pantalla de un ordenador posiblemente lleno de digitales ?.
No parece nada descabellado que los grandes criminales, esos que producen y a los que llegan grandes cantidades de dinero negro, estén
planeando un ataque como el de el 21/10 contra los bancos.
Todo el dinero sería negro o, lo que es lo mismo, todo el dinero podría empezar a ser blanco con tal de que se muestre en billetes.
Podría seguir se el dibujo apocalíptico con lo que se derivaría de ataques a los centros de gestión de el agua, de el fluido eléctrico, de las
distintas energías, de el tráfico y, en general, de todo lo que fluye y circula y nos hace vivir.
```

Figura 21 Resumen limpio.

Una vez generado el resumen y ordenado mediante el identificador de cada frase, se procesa el texto para quitar los identificadores y obtener un texto limpio para su lectura como se muestra en la figura 21.

Finalmente mostramos un ejemplo de resumen generado por el algoritmo Stonv al 30% del documento original.

¿Están preparadas nuestras instituciones básicas ante un ciberataque?

El pasado viernes 21 de octubre distintas empresas de tanto relieve como Spotify, NYT o Twitter sufrieron un ataque que las dejó paralizadas. Las explicaciones que ofrecieron del asunto distintos medios de comunicación y, sobre todo, distintos expertos eran para ponernos los pelos de punta. Incluso aunque no las entendiéramos -o tal vez por ello- estaban lejos de tranquilizarnos. Parecía que la Tercera Guerra Mundial había empezado; pero a través de nuestros cables sujetos a la red y nuestros aparatos más próximos: el ordenador, la impresora y cosas así. Se hablaba de que sólo había sido una especie de prueba, de entrenamiento, y que el próximo podría ir contra instituciones como hospitales o bancos. A partir de ahí, la imaginación se dirigía hacia el apocalipsis. Hasta ahora, la mayor parte del imaginario apocalíptico del cine o la literatura ponía como catástrofe originaria un desgraciado accidente nuclear, un ataque extraterrestre o una gran avería climática.
Pocas veces era el producto de un ataque cibernético en el que los resultantes del mismo habían sido producto de unos digitales que, armados con afán destructivo infinito, habían dormido en nuestros electrodomésticos hasta que alguien -no se sabe quién- los despertó. Sí será el apocalipsis si logran alcanzar instituciones como, por ejemplo, la Seguridad Social o Hacienda. Podrían desaparecer los datos y nadie cobraría su pensión, porque habría desaparecido de la documentación. Cada uno tendría que demostrar todo lo que había cotizado durante todos sus años de vida laboral. ¡Casi nada! Desde luego que se solucionaba el problema del acelerado adelgazamiento de la caja de las pensiones; pero de la peor forma posible.

Figura 22 Texto fuente.

En la figura 22 se muestra el texto fuente, en el cual esta remarcado con negrita el texto que forma parte del resumen generado por el algoritmo Stonv.

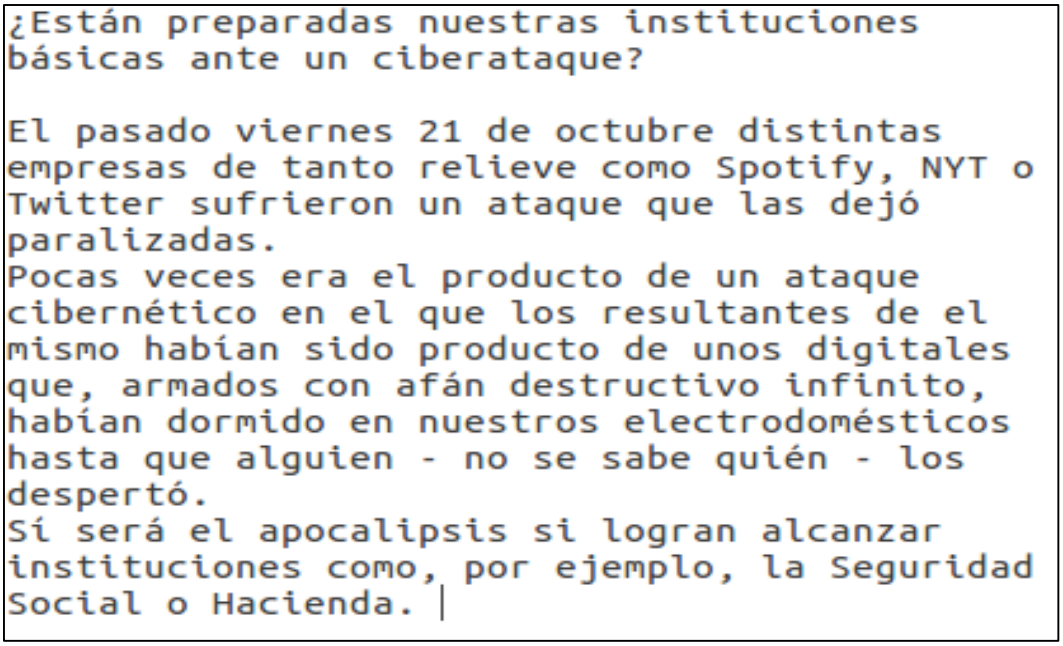


Figura 23 Resumen generado por Stonv.

En la figura 23 se muestra el resumen generado al 30% del texto original realizado por el algoritmo de resumen automático de texto Stonv.

6.3 Implementación de resumen y traducción automática

Para la implementación del algoritmo de resumen automático de texto Stonv y el sistema de traducción automática de Google se creó un sistema web como se muestra en la siguiente figura:

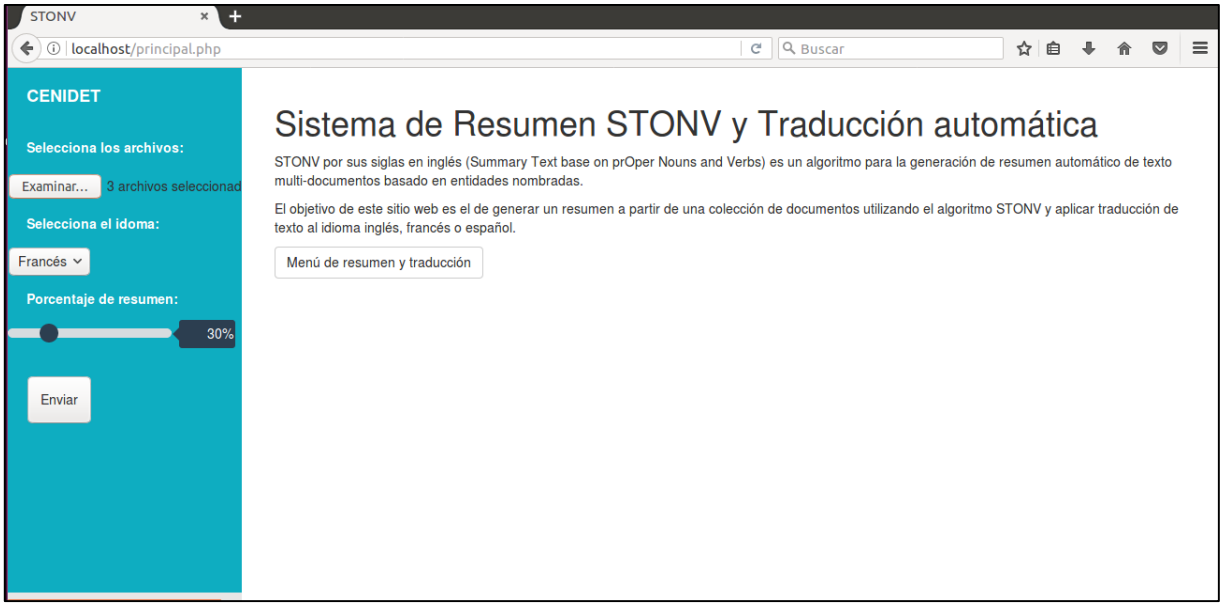


Figura 24 Sitio web de resumen y traducción automática de textos.

En la figura 24 se muestra el sitio web en donde tenemos resumen y traducción implementados en una sola herramienta, en la página tenemos los siguientes botones:

- Botón “Examinar”: Ingresaremos los documentos de texto que deseamos resumir
- Botón de idioma: Nos da 3 opciones a elegir para traducir el texto, español, inglés y francés, también tiene la opción de “Original” en donde no realiza ninguna traducción.
- Barra deslizadora: Nos permite elegir el porcentaje de resumen que deseamos obtener de los documentos fuente, dándonos una escala del 10 al 100% de resumen.

Una vez realizando lo anterior tenemos la opción de “Enviar” para realizar el resumen y la traducción de idioma seleccionados.

Para la creación de esta página se utilizaron las siguientes herramientas:

- HTML
- PHP
- Java Script
- CSS
- Bootstrap
- Apache Server

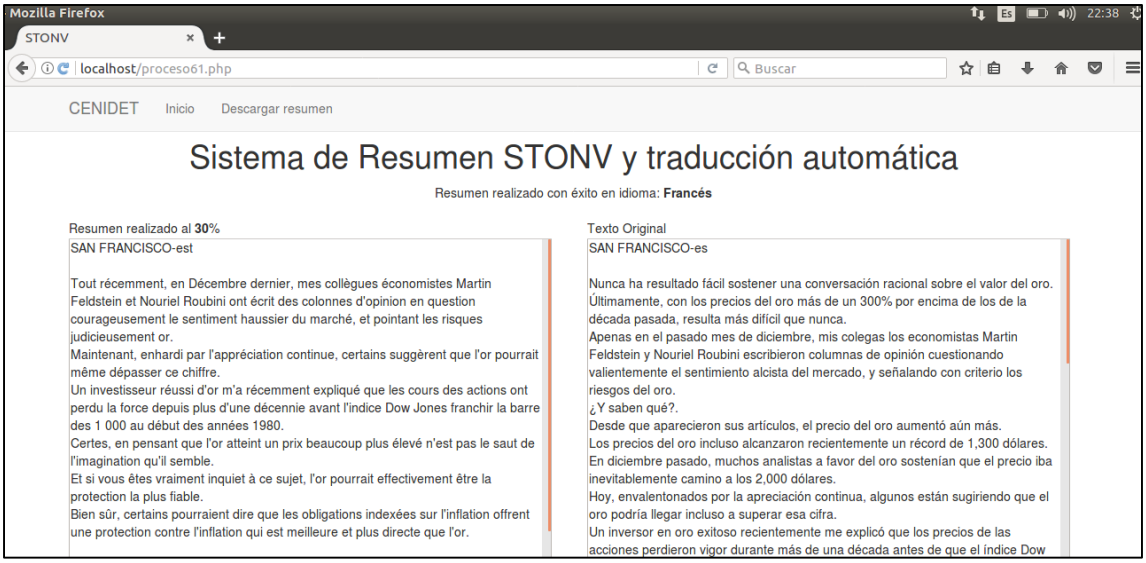


Figura 25 Resultado de resumen y traducción realizados.

En la figura 25 se muestra el resultado del resumen realizado al 30% de los documentos fuente y la traducción en idioma francés. La página nos muestra el resumen realizado y también nos muestra el texto original que hemos introducido.

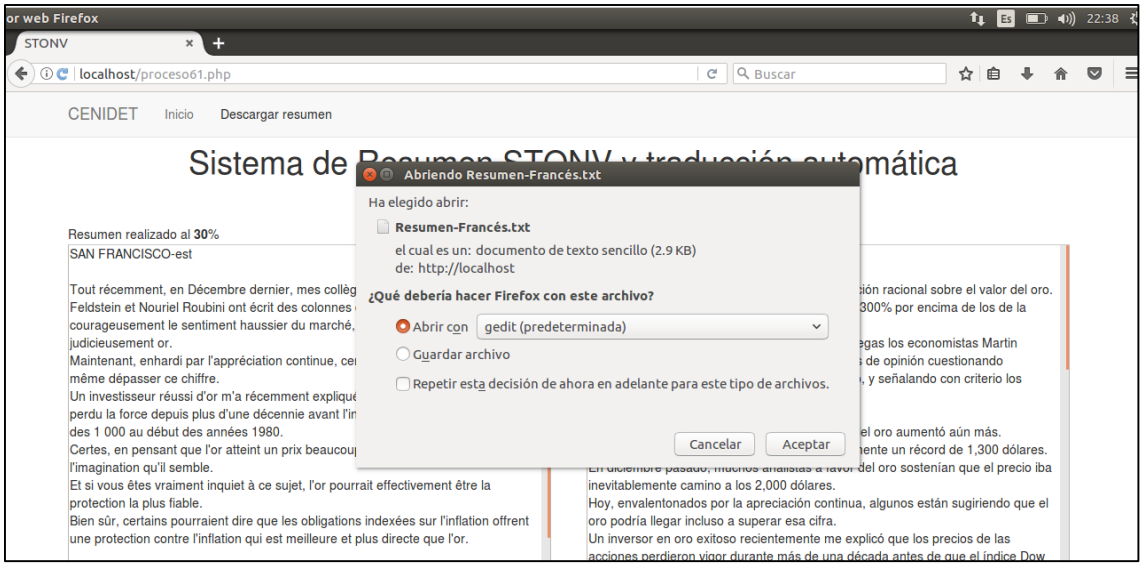


Figura 26 Descarga de resumen.

En la parte superior tenemos la opción de “inicio” que nos regresa a la página anterior que se mostró en la figura 24 y también tenemos la opción para descargar el resumen generado en “Descargar resumen” como se muestra en la figura 26.

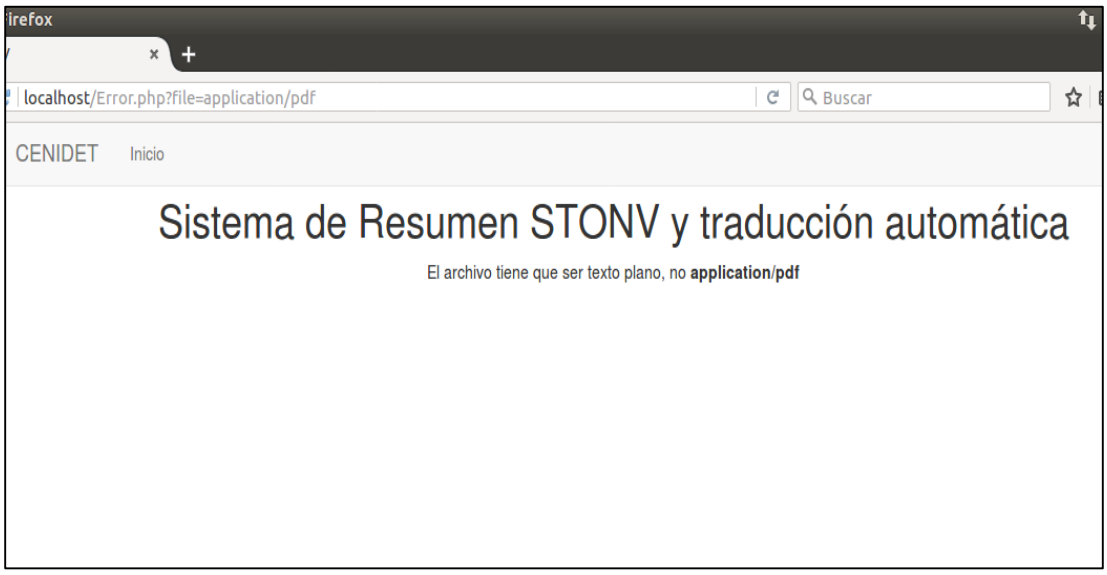


Figura 27 Ejecución de archivos en formatos diferentes a txt

El sistema de resumen Stonv y traducción automática solo admiten documentos de texto plano, es decir, formato txt, en caso de ingresar algún otro formato como pdf, Word, etc. La página nos mostrará el mensaje que se muestra en la figura 27 “El archivo tiene que ser texto plano, no **application/pdf**”.

Capítulo 7
Conclusiones y Trabajos Futuros

7.1 Conclusión

En este trabajo de tesis se logró crear una herramienta en la cual se realiza el resumen y traducción de documentos en un solo paso, permitiéndonos elegir el porcentaje de resumen que queremos obtener del documento fuente y de elegir entre los idiomas español, inglés y francés para traducir el resumen generado, obteniendo de esta forma una herramienta de resumen y traducción multi-documentos y multi idiomas. Así mismo se demostró que se obtienen mejores resultados en un resumen automático de texto con los tres algoritmos creados (Stonv, algoritmo Híbrido y Nasve), los cuales mediante las evaluaciones que se realizaron con la métrica Fresa resultaron mejores que los algoritmos *Artex, Cortex, Enertex Disicosum, Yate, Swesum, Open, Text summarizer, Pertinence summarizer y Word summarizer*. El algoritmo utilizado en esta herramienta es Stonv el cual es capaz de procesar texto en 13 idiomas diferentes, esto lo logra con base al etiquetado de Freeling, ya que procesa los siguientes idiomas: austriaco, catalán, inglés, francés, alemán, gallego, italiano, noruego, portugués, español, ruso, esloveno y galés, esto hace que Stonv sea un algoritmo de resumen de texto multi-idiomas.

El algoritmo Stonv obtuvo una precisión del 90% en su desempeño.

También se demostró que se obtienen mejores resultados realizando primero resumen y después la traducción, esta afirmación se respalda cuando se trabaja con algoritmos de resumen de tipo extractivos, ya que el texto fuente no se modifica al momento de realizar el resumen y el texto conserva su originalidad.

Se comprobó que el traductor *Google* entrega mejores resultados en sus traducciones mediante la métrica de evaluación Fresa, comparándolo con los sistemas de traducción *Systran* y *Reverso* posicionándolo en los mejores sistemas de traducción automática en la actualidad.

Por último, demostró que los algoritmos de resumen automático Híbrido y *Artex* obtienen mejores resultados que Stonv cuando se trabaja con textos cortos de 50 líneas o menos.

7.2 Trabajos futuros

Como trabajo futuro para el seguimiento de esta tesis se sugiere trabajar en la extracción de información de documentos en formatos pdf, Word etc. para poder realizar el resumen y la traducción de los mismos.
También se sugiere trabajar con el algoritmo Stonv para mejorar su precisión que es del 90%.

Trabajar con sistemas de traducción automática offline como Giza++ y Apertium para combinarlos con el algoritmo Stonv.

Realizar mejoras al sitio web creado y trabajarlo mediante un servidor en internet para brindar servicio público.

Referencias

[1] J.-M. Torres-Moreno, “Artex is AnotheR TEXt summarizer,” 2012.

[2] C. Armentano-Oller and M. L. Forcada, “Reutilización de datos lingüísticos para la creación de un sistema de traducción automática para un nuevo par de lenguas Re-use of linguistic data to create a machine translation system for a new language pair.”

[3] J. Antonio and C. Moya, “Sistema resumidor-traductor automático,” 2012.

[4] “Procesamiento de lenguajes naturales - Wikipedia, la enciclopedia libre.” [Online]. Available: https://es.wikipedia.org/wiki/Procesamiento_de_lenguajes_naturales. [Accessed: 13-Sep-2017].

[5] “Concepto de resumen - Definición en DeConceptos.com.” [Online]. Available: https://deconceptos.com/general/resumen. [Accessed: 13-Sep-2017].

[6] B. G. Chapero, “Resumen automático IA,” p. 19, 2009.

[7] “¿Qué es la traducción automática? | SYSTRAN – Tecnologías de traducción.” [Online]. Available: http://www.systran.es/systran/tecnologia-de-traduccion/que-es-la-traduccion-automatica/. [Accessed: 13-Sep-2017].

[8] “Welcome | FreeLing Home Page.” [Online]. Available: http://nlp.lsi.upc.edu/freeling/node/1. [Accessed: 08-May-2017].

[9] J.-M. Torres-Moreno, “FRESA,” *FRESA FRamework for Evaluating Summaries Automatically*, 2010. [Online]. Available: http://fresa.talne.eu/.

[10] M. González, “Estudio comparativo de traductores automáticos en línea: Systran, Reverso y Google,” *Núcleo*, vol. 22, no. 27, pp. 187–216, 2010.

[11] I. Da Cunha, J. M. Torres-Moreno, and P. Velazquez-, “Un algoritmo lingüístico-estadístico para resumen automático de textos especializados,” *Linguamática*, vol. 2, pp. 67–79, 2009.

[12] A. Molina, I. da Cunha, J.-M. Torres-Moreno, and P. Velazquez-Morales, “La compresión de frases: un recurso para la optimización de resumen automático de documentos,” *Linguamática*, vol. 2, no. 3, pp. 13–27, 2010.

[13] C. Armentano-Oller *et al.*, “Apertium, una plataforma de código abierto para el desarrollo de sistemas de traducción automática.”

[14] A. Mayor and I. Alegria, “Evaluaci´on de un sistema de traducci´on autom´atica basado en reglas o por qu´e BLEU s´olo sirve para lo que sirve.”

[15] D. K. Evans and J. L. Klavans, “Columbia Newsblaster: Multilingual News Summarization on the Web.”

[16] “Es Ciencia Y Tecnologia | El Huffington Post.” [Online]. Available: http://www.huffingtonpost.es/news/es-ciencia-y-tecnologia/. [Accessed: 08-May-2017].

[17] “Noticias de tecnología 2017 | Últimas noticias tecnológicas | EiTB Tecnología.” [Online]. Available: http://www.eitb.eus/es/noticias/tecnologia/. [Accessed: 08-May-2017].

[18] “Google Noticias.” [Online]. Available: https://news.google.com.mx/. [Accessed: 08-May-2017].

[19] “Moses - Paralelo-Corpus.” [Online]. Available: http://www.statmt.org/moses/?n=moses.baseline. [Accessed: 13-Sep-2017].

[20] C. Y. Lin, “Rouge: A package for automatic evaluation of summaries,” *Proc. Work. text Summ. branches out (WAS 2004)*, no. 1, pp. 25–26, 2004.

[21] J. C. Baez and T. Fritz, “A Bayesian Characterization of Relative Entropy,” Feb. 2014.

[22] “Real Academia Española.” [Online]. Available: http://www.rae.es/. [Accessed: 31-May-2017].

Anexos

Anexo 1

En anexo 1 se muestra las evaluaciones realizadas en el artículo "Estudio comparativo de traductores automáticos en línea: Systran, Reverso y Google" [10]

Tabla 9 Ejemplos de traducción de homonimia y polisemia

TIPO	FRASE/ORACIÓN ORIGINAL	SYSTRAN	GOOGLE	REVERSO	TRADUCCIÓN MANUAL
Ambigüedad léxica: polisemia	1. light - bill - This is a light suitcase. - The light bill is here.	Luz cuenta - Esto es una maleta ligera. - La cuenta ligera está aquí.	Ligero/luz cuenta - Se trata de una maleta de luz. - El proyecto de ley de luz está aquí.	Luz Cuenta (proyecto de la ley) - Esto es una maleta ligera (de luz). - La cuenta (El proyecto de la ley) ligera (de luz) está aquí.	Luz/ ligero/a Cuenta/proyecto de ley - La maleta es ligera. - La factura de la luz está aquí.
	2. To hit - Somebody hit him at school. - The car hit the tree. - The police fired and hit their target.	Para golpear - Alguien lo golpeó en la escuela. - El coche golpeó el árbol. - La policía encendida y golpeada su blanco.	Golpear - Alguien lo golpeó en la escuela. - El auto chocó contra el árbol. - La policía disparó y golpeó a su objetivo.	Golpear - Alguien lo golpeó en la escuela. - El coche golpea el árbol. - La policía encendió (despidió) y dado en su blanco.	Golpear/chocar/ dar en el blanco - Alguien le pegó en la escuela. - El coche chocó contra el árbol. - La policía disparó y dio en el blanco.
Ambigüedad léxica: homonimia	3. Box - He gave me a box. - He gave him a box to the face.	- Él me dio una caja. - Él le dio una caja a la cara.	- Me dio una caja. - Le dio una caja a la cara.	- Él me dio una caja. - Él le dio una caja a la cara.	- Me dio una caja. - Le dio una bofetada.
	4. Match - I went to a match. - They are a perfect match.	- Fui a un fósforo. - Son fósforo perfecto.	- Fui a un partido. - Son pareja perfecta.	- Fui a un fósforo (partido). - Ellos son el partido perfecto.	- Fui a un partido. - Hacen una pareja perfecta.

Tabla 10 Ejemplos de traducción de calcos léxicos y sintácticos

TIPO	FRASE/ORACIÓN ORIGINAL	SYSTRAN	GOOGLE	REVERSO	TRADUCCIÓN MANUAL
Fraseología. Calco léxico.	5. I endorse your candidacy.	Endoso su candidatura.	Apoyo su candidatura.	Endoso (Apruebo) a su candidatura.	Apoyo tu candidatura.
Fraseología. Calco sintáctico. Orden oracional.	6. If your boss calls you.	Si su jefe les llama.	Si tu jefe te llama.	Si su jefe le llama.	Si te llama tu jefe.
	7. He spends the entire day playing with the computer.	Él pasa el día entero que juega con el ordenador.	Se pasa todo el día jugando con el ordenador.	Él gasta (pasa) el día entero jugando con el ordenador.	Se pasa todo el día/ el día entero jugando con el ordenador.
Fraseología. Calco sintáctico. Sintagma verbal.	8. This looks to be a very good decision.	Esto mira para ser una decisión muy buena.	Esto parece ser una muy buena decisión.	Esto mira para ser una decisión muy buena.	Parece una muy buena decisión.
Fraseología. Calco sintáctico. Sintagma nominal.	9. Take your time.	Tarden su tiempo.	Tómese su tiempo.	Tome su tiempo.	Tómate tiempo.
Fraseología. Calco sintáctico. Sintagma preposicional.	10. I am taking the dog with me.	Estoy tomando el perro conmigo.	Estoy tomando el perro conmigo.	Tomo el perro conmigo.	Me llevo al perro.

Tabla 11 Ejemplos de traducción de construcciones pasivas

TIPO	FRASE/ORACIÓN ORIGINAL	SYSTRAN	GOOGLE	REVERSO	TRADUCCIÓN MANUAL
Pasiva	11. He was expelled from school by the principal.	El principal lo expulsó de escuela.	Fue expulsado de la escuela por el director.	Él fue expulsado de la escuela por el principal.	El director lo expulsó de la escuela.
	12. Many things were said at that meeting.	Muchas cosas fueron dichas en esa reunión.	Muchas cosas se dijeron en esa reunión.	Muchas cosas fueron dichas en aquella reunión.	Se dijeron muchas cosas en aquella reunión.
	13. Tell me Paul, where were you born?	¿Me dice Paul dónde nació usted?	Dime Pablo, ¿dónde nació usted?	¿Dígame dónde fue nacido usted?	Dime Pablo, ¿dónde naciste?
	14. A lot of money was spent.	Mucho dinero estuvo pasado.	Una gran cantidad de dinero se gastó.	Mucho dinero fue gastado (pasado).	Se gastó mucho dinero.
	15. Mr. Roberts was taken to hospital.	Llevaron Sr. Roberts al hospital.	El Sr. Roberts fue llevado al hospital.	Sr. Roberts fue tomado al hospital.	Al Sr. Roberts le llevaron al hospital.

Tabla 12 Ejemplos de traducción de refranes

TIPO	FRASE/ORACIÓN ORIGINAL	SYSTRAN	GOOGLE	REVERSO	TRADUCCIÓN MANUAL
Refranes	16. To keep a stiff upper lip.	Para mantener un labio superior tieso.	Guardar (Mantener) un labio tieso superior.	Guardar (mantener) un labio tieso superior.	A mal tiempo, buena cara.
	17. It is the early bird that catches the worm.	Es el ave temprana que las capturas que el gusano.	Esto es el temprano pájaro que coge el gusano.	Esto es el temprano pájaro que coge el gusano.	A quien madruga, Dios le ayuda.
	18. Like father, like son.	De tal palo tal astilla.	De tal palo tal astilla.	Como padre, como hijo.	De tal palo tal astilla.
	19. The apple doesn't fall far from the tree.	La manzana no cae lejos del árbol.	La manzana no cae lejos del árbol.	La manzana no se cae lejos del árbol.	De tal palo tal astilla.
	20. He's a chip off the old block.	Él es un microprocesa-dor del viejo bloque.	Él es una astilla del viejo tronco.	Él es una viruta (un chip) del viejo bloque	De tal palo tal astilla
	21. The shoemaker's son goes always barefoot.	El hijo del zapatero va siempre descalzo.	El hijo del zapatero va siempre descalza.	El hijo del zapatero va siempre con los pies desnudos.	En casa del herrero, cuchara de palo.
	22. A bird in the hand is worth two in the bush.	Un pájaro en la mano vale dos en el arbusto.	Más vale pájaro en ma-no que ciento volando.	Más vale pájaro en mano que cien volando.	Más vale pájaro en ma-no que ciento volando.
	23. It's the last straw that breaks the camel's back.	Es la última gota que rompe el camello detrás.	Es la gota que colma el vaso.	Esto es la gota que desbordó el vaso que rompe el camello atrás.	Es la gota que colma el vaso.

Tabla 13 Ejemplos de traducción de modismos

TIPO	FRASE/ORACIÓN ORIGINAL	SYSTRAN	GOOGLE	REVERSO	TRADUCCIÓN MANUAL
Modismos	24. Add fuel to the fire.	Añadan el combustible al fuego.	Añadir leña al fuego.	Echar leña al fuego.	Echar leña al fuego.
	25. To hit the nail on the head.	Para golpear el clavo en la cabeza.	Para golpear el clavo en la cabeza.	Golpear la uña en la cabeza.	Dar en el clavo.
	26. To be hand in glove with somebody.	Para ser mano en guante con alguien.	Para ser la mano con alguien.	Ser de la mano en guante con alguien.	Ser uña y carne.
	27. As good as it gets.	Tan bueno como él consigue.	Tan bueno como se pone.	Tan bueno como se pone.	Mejor imposible.
	28. To get cold feet.	Para conseguir pies fríos.	Para obtener pies fríos.	Hacer frío pies.	Echarse para atrás.
	29. As straight as a die.	Tan derecho como un dado.	Tan recto como un dado.	Derecho como una vela.	Más recto que una vela.

Tabla 14 Ejemplos de traducción adverbiales

TIPO	FRASE/ORACIÓN ORIGINAL	SYSTRAN	GOOGLE	REVERSO	TRADUCCIÓN MANUAL
Locución adverbial	30. I met him sometime ago.	Lo encontré hace algún tiempo.	Lo conocí hace algún tiempo.	Lo encontré algún día.	Lo conocí hace tiempo.
	31. I go to visit her every other day.	Voy a visitarla cada otro día.	Voy a visitar todos los días de otros.	Voy a visitarla cada dos días.	Voy a verla un día sí y otro no.
	32. I see him less and less.	Lo veo cada vez menos.	Yo lo veo cada vez menos.	Yo lo veo menos y menos.	Lo veo cada vez menos.
	33. We will get the money somehow or other.	Conseguiremos el dinero de alguna manera u otro.	Vamos a sacar el dinero de alguna manera o de otra índole.	Conseguiremos el dinero de uno o de otro modo.	Conseguiremos el dinero como sea.

Tabla 15 Ejemplos de traducción de colocaciones

TIPO	FRASE/ORACIÓN ORIGINAL	SYSTRAN	GOOGLE	REVERSO	TRADUCCIÓN MANUAL
Locución adverbial	30. I met him sometime ago.	Lo encontré hace algún tiempo.	Lo conocí hace algún tiempo.	Lo encontré algún día.	Lo conocí hace tiempo.
	31. I go to visit her every other day.	Voy a visitarla cada otro día.	Voy a visitar todos los días de otros.	Voy a visitarla cada dos días.	Voy a verla un día sí y otro no.
	32. I see him less and less.	Lo veo cada vez menos.	Yo lo veo cada vez menos.	Yo lo veo menos y menos.	Lo veo cada vez menos.
	33. We will get the money somehow or other.	Conseguiremos el dinero de alguna manera u otro.	Vamos a sacar el dinero de alguna manera o de otra índole.	Conseguiremos el dinero de uno o de otro modo.	Conseguiremos el dinero como sea.

Tabla 16 Ejemplos de traducción de siglas

TIPO	SIGLA	SYSTRAN	GOOGLE	REVERSO	TRADUCCIÓN MANUAL
Siglas	40. EU	UE	De la UE	Unión Europea	UE, Unión Europea
	41. UN	LA O.N.U	De las Naciones Unidas	Naciones Unidas	ONU, Organización de las Naciones Unidas
	42. CARICOM	CARICOM	CARICOM	CARICOM	Comunidad Caribeña

Tabla 17 Ejemplos de traducción de nombres propios

TIPO	NOMBRE	SYSTRAN	GOOGLE	REVERSO	TRADUCCIÓN MANUAL
Nombres propios	43. John White	Blanco de Juan	John White	John White	John/Juan White
	44. Mary Carpenter	Carpintero de Maria	Mary Carpenter	María Carpenter	Mary/María Carpenter

Tabla 18 Ejemplos de traducción de topónimos

TIPO	TOPÓNIMO	SYSTRAN	GOOGLE	REVERSO	TRADUCCIÓN MANUAL
Topónimos	**45. King Street**	Rey Street	King Street	Calle de Rey	King Street/Calle del Rey
	46. Old Hope Road	Camino viejo de la esperanza	Old Hope Road	Viejo Camino de Esperanza	Old Hope Road
	47. Cape Town	Cape Town	Ciudad del Cabo	Ciudad del Cabo	Ciudad del Cabo
	48. Poland	Polonia	Polonia	Polonia	Polonia
	49. The Netherlands	Los Países Bajos	Los Países Bajos	Países Bajos	Países Bajos
	50. Cayman Islands	Islas Caimán	Islas Caimán	Islas Caimán	Islas Caimán

Anexo 2

En anexo 2 se muestra los valores obtenidos en las evaluaciones realizados a los algoritmos de resumen automáticos con textos cortos de 50, 40 y 30 líneas

Tabla 19 Evaluación de resúmenes de un texto de 50 líneas.

Evaluación de resúmenes de un texto de 50 líneas				
	*** Moyennes|Average / 1 eval ***			

SYSTEME	FRESA_4	FRESA_2	FRESA_1	FRESA_M
Hibrido	**0,30808**	**0,31443**	**0,45831**	**0,36027**
Stonv	0,28902	0,29196	0,44854	0,34323
Nasve	0,28803	0,29094	0,44759	0,34219
Artex	0,24671	0,254	0,37441	0,29171
Nason	0,20367	0,20778	0,371	0,26082
BaseLine1	0,19129	0,19689	0,35565	0,24794
LastBaseLine	0,20221	0,20552	0,32673	0,24482
BaseLine2	0,12597	0,13729	0,29273	0,18533
BaseLine3	0,11281	0,12392	0,24681	0,16118
FirstBaseLine	0,10237	0,10106	0,19152	0,13165

Tabla 20 Evaluación de resúmenes de un texto de 40 líneas.

Evaluación de resúmenes de un texto de 40 líneas				
	*** Moyennes|Average / 1 eval ***			

SYSTEME	FRESA_4	FRESA_2	FRESA_1	FRESA_M
Nasve	**0,21712**	**0,23064**	**0,37548**	**0,27441**
Stonv	**0,21712**	**0,23064**	**0,37548**	**0,27441**
Hibrido	0,21815	0,22801	0,34541	0,26386
BaseLine2	0,19151	0,19042	0,36059	0,24751
Artex	0,18064	0,18803	0,2791	0,21592
BaseLine3	0,16557	0,17801	0,30027	0,21462
Nason	0,17053	0,17668	0,29272	0,21331
BaseLine1	0,12934	0,13544	0,26216	0,17565
LastBaseLine	0,14209	0,14077	0,2388	0,17389
FirstBaseLine	0,11641	0,11465	0,16706	0,13271

Tabla 21 Evaluación de resúmenes de un texto de 30 líneas.

Evaluación de resúmenes de un texto de 30 líneas				
	*** Moyennes|Average / 1 eval ***			

SYSTEME	FRESA_4	FRESA_2	FRESA_1	FRESA_M
Artex	**0,23636**	**0,2446**	**0,33592**	**0,27229**
Stonv	0,21241	0,22143	0,30461	0,24615
Nasve	0,21241	0,22143	0,30461	0,24615
Hibrido	0,21241	0,22143	0,30461	0,24615
LastBaseLine	0,12522	0,12454	0,24648	0,16541
BaseLine3	0,11905	0,12183	0,25264	0,16451
BaseLine2	0,10922	0,11134	0,24523	0,15526
BaseLine1	0,09309	0,10406	0,20318	0,13345
Nason	0,09256	0,10522	0,19996	0,13258
FirstBaseLine	0,10054	0,10814	0,16347	0,12405

Anexo 3

En el anexo 3 se muestran las 100 frases que se utilizaron para evaluar la precisión, el recall y f-measure del algoritmo para la generación de resumen automático Stonv basado en entidades nombradas, en donde:

NP = Nombres propios,

(FP)NP = falsos negativos de nombres propios,

V = verbos

(FN)V = Falsos negativos de verbos

Tabla 22 Frases para la evaluación del algoritmo Stonv.

No.	Frase		NP	(FP) NP	V	(FN)V
1	buenas noticias y demasiadas excusas.	REAL	0		0	
		PREDICHO	0		0	
2	starmus, una patada en España y acaba en Noruega.	REAL	3		1	
		PREDICHO	2		1	
3	¿están preparadas nuestras instituciones básicas ante un ciberataque?	REAL	0		3	
		PREDICHO	0		2	
4	el alto precio de cerrar la brecha digital en el mundo	REAL	0		1	
		PREDICHO	0		1	
5	el principal reto en la lucha por el clima sigue residiendo, desde mi punto de vista, en la dificultad de cambiar actitudes individuales. ~	REAL	0		4	
		PREDICHO	0		3	
6	quizá no merezca la pena detenerse a convencer a unos pocos de que el cambio climático no es un complot internacional contra el ciudadano de a pie. ~	REAL	0		4	
		PREDICHO	0		4	

7	más bien, hay que centrarse en romper con la mentalidad del "yo ya separo mi basura, pero esto lo tienen que resolver los políticos" (entre otras cosas porque, como explico arriba, los políticos ya están en ello, aunque no sea noticia de portada). ~	REAL	0		9	
		PREDICHO	0		9	
8	los ciudadanos tenemos mucho peso en la protección del medio ambiente; por ejemplo, en el ámbito de la reducción y reutilización de residuos (ya habréis visto la cantidad de). ~	REAL	0		5	
		PREDICHO	0		3	
9	desde luego que, como ya he argumentado repetidas veces en este blog, los políticos pueden, mediante la regulación, influir mucho en las decisiones que tomamos (recientemente, por ejemplo, él ha introducido ventajas fiscales sobre reparaciones con este fin). ~	REAL	0		11	
		PREDICHO	0		8	
10	pero tenemos que entender que vivir en democracia no es un asunto de 20 minutos cada 4 años. ~	REAL	0		4	
		PREDICHO	0		4	
11	es momento de que los ciudadanos admitamos que es responsabilidad de todos (incluido uno mismo) cumplir con los compromisos establecidos en el Acuerdo de París, y reflexionar sobre el impacto de nuestras rutinas sobre el bienestar de las generaciones futuras. ~	REAL	1		7	
		PREDICHO	1		7	
12	para esto, quizá a alguno le sea de utilidad el libro o seguir en Instagram a Basta de excusas. ~	REAL	1		2	
		PREDICHO	1	1	2	

13	no encuentro palabras para definir lo que ha ocurrido con el Festival Starmus en España. ~	REAL	2		4	
		PREDICHO	2		4	
14	no las encuentro porque me produce una tremenda tristeza que este país haya dado una patada a un evento de estas características, tan potente que lo ha lanzado hasta Noruega. ~	REAL	1		6	
		PREDICHO	1		5	
15	allí, ha caído en blando, sobre unos brazos abiertos, porque allí no tienen políticos ni empresas tan cerriles como las que tenemos aquí. ~	REAL	0		4	
		PREDICHO	0		4	1
16	pero no sólo maltratamos, es que encima culpamos al organizador, el astrofísico Garik Israelian, de que se fiara de los compromisos contractuales, de que pensara que no le iban a dejar en la estacada. ~	REAL	1		6	
		PREDICHO	1		6	1
17	no voy a recordar todo lo que ya conté un mes antes del evento, celebrado a finales de junio, cuando quedó patente que de nuevo algo no iba bien a nivel financiero con Starmus durante la rueda de prensa de presentación. ~	REAL	1		7	
		PREDICHO	1		7	
18	lo único que he sabido después es que fue un éxito de afluencia, que todos los asistentes quedaron encantados y que el Cabildo de Tenerife, presidido por Carlos Alonso, no consiguió los patrocinios de algunas grandes empresas por 350. ~	REAL	2		8	
		PREDICHO	2		8	
19	1 000 euros, a lo que se había comprometido y que no ha negado. ~	REAL	0		4	
		PREDICHO	0		4	

20	en concreto, me refiero al de las grandes empresas ITER y ATOS, cuyos responsables sí corrieron a hacerse con el científico Stephen Hawking. ~	REAL	3		3	
		PREDICHO	3		3	
21	me sorprende que prácticamente ningún medio de comunicación español se preocupara, salvo de pasada, de unos problemas financieros conocidos desde entonces y que ponían en riesgo su continuidad, al menos en España; que en sus tres ediciones no tuviera apoyo del Gobierno central (salvo también para ir a hacerse la foto con grandes genios del planeta); y que ninguna de las grandes empresas de este país quisiera ligar su marca a un encuentro de este nivel, que no ha dejado de crecer y pasará a la historia de la divulgación científica.~	REAL	1		13	
		PREDICHO	1	1	12	2
22	guste a unos su formato o no guste. ~	REAL	0		2	
		PREDICHO	0		2	
23	parece que a los noruegos sí les parece estupendo. ~	REAL	0		1	1
		PREDICHO	0		1	1
24	desde luego, me consta que ninguno de los grandes empresarios del IBEX35 podrá decir que no se les contactó, pero unos quisieron aprovecharse de la presencia de algunos famosos (como Hawking) sin poner un euro y otros ni se molestaron en abrir la puerta, en contestar a una carta, a una llamada. ~	REAL	2		10	
		PREDICHO	1		10	
25	así de patético es este país. ~	REAL	0		1	
		PREDICHO	0		1	
26	así de triste. ~	REAL	0		0	
		PREDICHO	0		0	

27	leo en algunos medios que el Cabildo de Tenerife dice en torno al 80% de la financiación de Starmus ha sido pública. ~	REAL	2		4	
		PREDICHO	2		3	
28	también leo (no en los medios) que Starmus lo niega () y que más del 50% se ha financiado con las entradas. ~	REAL	1		4	
		PREDICHO	1		3	
29	leo en los medios que resulta que tiene deudas en Canarias por pagar. ~	REAL	1		4	
		PREDICHO	1		3	
30	pero vamos a ver, si alguien se compromete (y no de palabra, que se las lleva el viento) a hacer algo que luego no hace (como quedó claro en la presentación) y eso afecta al balance, final ¿de quién es la culpa? ~	REAL	0		10	
		PREDICHO	0		9	
31	desde luego, así se hunden iniciativas en este país de pandereta. ~	REAL	0		1	
		PREDICHO	0		1	
32	con todo lo anterior, no me sorprende que el Festival Starmus tuviera otras más cariñosas. ~	REAL	1		2	
		PREDICHO	1		2	
33	más de una, según tengo entendido. ~	REAL	0		1	
		PREDICHO	0		1	
34	no sólo por la resonancia mundial del evento (incluyendo páginas en él y o BBC, por mencionar algunas publicaciones), sino porque eran muchos los jóvenes (mayoritariamente canarios) que esperaban cada año la oportunidad de acercarse a la ciencia de más alto nivel durante una semana. ~	REAL	1		5	
		PREDICHO	1		5	

35	los canarios, por 350 euros, y los demás por 500 (si se sacaban con antelación). ~	REAL	0		1	
		PREDICHO	0		1	
36	si hubieran costado 1. ~	REAL	0		2	
		PREDICHO	0		2	
37	seguramente se habría financiado sin ese apoyo público. ~	REAL	0		2	
		PREDICHO	0		2	
38	pero en ese caso, ¿qué jóvenes hubieran ido? ~	REAL	0		2	
		PREDICHO	0		2	
39	por otro lado, ¿alguien en España tiene capacidad de movilizar a los mismos personajes que Garik Israelian? ~	REAL	2		2	
		PREDICHO	2		2	
40	como ciudadana que paga impuestos, prefiero que se financien estos eventos con dinero antes de que vaya para, mucho más millonarias, a la banca () o a s.~	REAL	0		4	
		PREDICHO	0		4	
41	eso por no hablar de la Iglesia. ~	REAL	1		1	
		PREDICHO	1		1	
42	puestos a elegir. ~	REAL	0		1	
		PREDICHO	0		1	
43	no sé si Starmus 2018 será o no en España, o en Canarias. ~	REAL	3		1	
		PREDICHO	3		1	1
44	pero tal como van las cosas, mucho me temo que no.~	REAL	0		1	
		PREDICHO	0		1	1
45	pero vamos, es sólo otra gran iniciativa española, de un emprendedor, que se va fuera. ~	REAL	1		3	
		PREDICHO	0		3	
46	con políticos y empresarios como los de ahora, Colón no habría descubierto América para el resto del mundo fijo. ~	REAL	2		2	
		PREDICHO	2		2	

47	el pasado viernes 21 de octubre distintas empresas de tanto relieve como Spotify, NYT o Twitter sufrieron un ataque que las dejó paralizadas. ~	REAL	3		2	
		PREDICHO	3		2	1
48	las explicaciones que ofrecieron del asunto distintos medios de comunicación y, sobre todo, distintos expertos eran para ponernos los pelos de punta. ~	REAL	0		3	
		PREDICHO	0		3	
49	incluso aunque no las entendiéramos -o tal vez por ello- estaban lejos de tranquilizarnos. ~	REAL	0		3	
		PREDICHO	0		3	
50	parecía que la Tercera Guerra Mundial había empezado; pero a través de nuestros cables sujetos a la red y nuestros aparatos más próximos: el ordenador, la impresora y cosas así. ~	REAL	0		2	
		PREDICHO	1		2	1
51	se hablaba de que sólo había sido una especie de prueba, de entrenamiento, y que el próximo podría ir contra instituciones como hospitales o bancos. ~	REAL	0		5	
		PREDICHO	0		5	
52	a partir de ahí, la imaginación se dirigía hacia el apocalipsis. ~	REAL	1		1	
		PREDICHO	0		1	
53	hasta ahora, la mayor parte del imaginario apocalíptico del cine o la literatura ponía como catástrofe originaria un desgraciado accidente nuclear, un ataque extraterrestre o una gran avería climática. ~	REAL	0		2	
		PREDICHO	0		1	1

54	pocas veces era el producto de un ataque cibernético en el que los resultantes del mismo habían sido producto de unos digitales que, armados con afán destructivo infinito, habían dormido en nuestros electrodomésticos hasta que alguien - no se sabe quién- los despertó. ~	REAL	1		8	
		PREDICHO	0		8	
55	sí será el apocalipsis si logran alcanzar instituciones como, por ejemplo, la Seguridad Social o Hacienda. ~	REAL	3		3	
		PREDICHO	2		3	
56	podrían desaparecer los datos y nadie cobraría su pensión, porque habría desaparecido de la documentación. ~	REAL	0		5	
		PREDICHO	0		5	
57	cada uno tendría que demostrar todo lo que había cotizado durante todos sus años de vida laboral. ~	REAL	0		4	
		PREDICHO	0		4	
58	¡Casi nada! Desde luego que se solucionaba el problema del acelerado adelgazamiento de la caja de las pensiones; pero de la peor forma posible. ~	REAL	0		3	
		PREDICHO	0		2	
59	lo mismo podría pasar con Hacienda, aun cuando aquí, bastantes podrían salir beneficiados. ~	REAL	1		4	
		PREDICHO	1		4	1
60	eso hablando de instituciones públicas que, al menos, parecen cibernéticamente seguras. ~	REAL	0		1	
		PREDICHO	0		1	1
61	al menos, eso espero. ~	REAL	0		1	
		PREDICHO	0		1	
62	pero pensemos en otras, como nuestras universidades. ~	REAL	0		1	
		PREDICHO	0		1	

63	supongamos que un ciberataque borra toda huella de los archivos universitarios: ¿cómo se podrá mostrar quien tiene el título para ejercer una profesión, si antes no ha tenido la ventura de conseguirlo en el viejo papel? ~	REAL	0		9	
		PREDICHO	0		9	
64	¿Tendrá que ir, profesor por profesor, solicitando que le recordara su paso por las aulas hace ya cinco, diez o quince años? ~	REAL	0		5	
		PREDICHO	0		4	1
65	todo lo anterior tal vez no sea suficiente para generar esos lúgubres escenarios catastrofistas; pero piensen en un ciberataque al sistema financiero, a cada uno o la mayor parte de los grandes bancos. ~	REAL	0		2	
		PREDICHO	0		2	1
66	un borrado de nuestros ahorros o de nuestras inversiones. ~	REAL	0		1	
		PREDICHO	0		1	
67	tal vez también de nuestras deudas; pero los bancos ya se han encargado de que nuestras deudas mayores, que son de origen hipotecario, queden bien registradas -negro sobre blanco- ante un notario. ~	REAL	0		5	
		PREDICHO	0		5	
68	incluso podría ocurrir que se borrase lo que hemos ido amortizando de la deuda, quedando sólo el registro del origen total de la misma, el acto en el que entregamos la garantía de nuestra casa. ~	REAL	0		8	
		PREDICHO	0		8	

69	¿No suena un poco extraño que las deudas que tenemos con los bancos quedan inscritas en papel y que las que tienen los bancos con nosotros, a partir de los depósitos, sólo estén sostenidas por una evanescente pantalla de un ordenador posiblemente lleno de digitales? ~	REAL	0		7	
		PREDICHO	0		7	
70	a partir de aquí, se recomienda imprimir periódicamente. ~	REAL	0		2	
		PREDICHO	0		2	
71	no parece nada descabellado que los grandes criminales, esos que producen y a los que llegan grandes cantidades de dinero negro, estén planeando un ataque como el del 21/10 contra los bancos. ~	REAL	0		6	
		PREDICHO	0		5	1
72	al fin y al cabo, qué es el dinero negro sino el dinero no registrado. ~	REAL	0		2	
		PREDICHO	0		2	
73	imposible identificar un billete de moneda circulante de dinero negro, de un billete de dinero blanco. ~	REAL	0		1	
		PREDICHO	0		1	
74	la diferencia es que el dinero blanco es el registrado y, por lo tanto, visible y del que puede seguirse el trayecto. ~	REAL	0		5	
		PREDICHO	0		5	
75	pero ¿qué ocurriría si desapareciera el dinero blanco por un borrado masivo de los registros bancarios? ~	REAL	0		3	
		PREDICHO	0		2	
76	todo el dinero sería negro o, lo que es lo mismo, todo el dinero podría empezar a ser blanco con tal de que se muestre en billetes. ~	REAL	0		6	
		PREDICHO	0		6	

77	podría seguirse el dibujo apocalíptico con lo que se derivaría de ataques a los centros de gestión del agua, del fluido eléctrico, de las distintas energías, del tráfico y, en general, de todo lo que fluye y circula y nos hace vivir. ~	REAL	0		9	
		PREDICHO	0		7	
78	la sociedad líquida, de la que nos habla Bauman, parece enormemente expuesta. ~	REAL	1			
		PREDICHO	1		3	
79	especialmente si la comparamos con las sociedades sólidas. ~	REAL	0		1	
		PREDICHO	0		1	
80	se trata sólo de escenarios que, supongo, los distintos responsables de seguridad habrán previsto y que, por lo tanto, no hay motivos para preocuparse. ~	REAL	0		5	
		PREDICHO	0		5	1
81	todo está controlado y sería mejor dejar todas estas paranoicas pesadillas. ~	REAL	0		3	
		PREDICHO	0		3	1
82	pero antes de la crisis bancaria también nos dijeron que teníamos uno de los sistemas financieros más sanos del mundo y cosas así. ~	REAL	0		2	
		PREDICHO	0		2	
83	la Organización de Naciones Unidas ha fijado un ambicioso conjunto de objetivos a alcanzar para el año 2030. ~	REAL	1		3	
		PREDICHO	1		3	
84	"La Agenda es un plan de acción en favor de las personas, el planeta y la prosperidad. ~	REAL	0		1	
		PREDICHO	0	1	1	
85	también tiene por objeto fortalecer la paz universal dentro de un concepto más amplio de la libertad". ~	REAL	0			

		PREDICHO	0		2	
86	¿Están preparadas nuestras instituciones básicas ante un ciberataque? ~	REAL	0		2	
		PREDICHO	0		2	
87	El pasado viernes 21 de octubre distintas empresas de tanto relieve como Spotify, NYT o Twitter sufrieron un ataque que las dejó paralizadas. ~	REAL	3		3	
		PREDICHO	3		3	
88	Las explicaciones que ofrecieron del asunto distintos medios de comunicación y, sobre todo, distintos expertos eran para ponernos los pelos de punta. ~	REAL	1		3	
		PREDICHO	1		3	
89	Incluso aunque no las entendiéramos -o tal vez por ello- estaban lejos de tranquilizarnos. ~	REAL	0		3	
		PREDICHO	0		3	
90	Mientras que las empresas juzgan que es sensiblemente inferior a lo que pensamos. ~	REAL	0		3	
		PREDICHO	0		3	
91	Sin embargo, parece que es suficiente para algunas empresas de telecomunicaciones de Estados Unidos como para proponer un modelo. ~	REAL	1		3	
		PREDICHO	1		3	
92	El personaje de Goethe intercambió su alma por el conocimiento sin límite y los placeres terrenales; la privacidad y la libertad de información es situado como precio del acceso a la Internet de banda ancha. ~	REAL	2		4	
		PREDICHO	2		3	
93	Hasta ahora, la mayor parte del imaginario apocalíptico del cine o la literatura ponía como catástrofe originaria un desgraciado accidente nuclear, un ataque extraterrestre o una gran avería climática. ~	REAL	0		2	
		PREDICHO	0		1	

94	Siguiendo el modelo generalizado por las plataformas digitales (Google, Facebook, etc.), dejar ser rastreado con fines comerciales proporcionaría descuentos en la factura mensual del servicio de acceso a Internet. ~	REAL	3		5	
		PREDICHO	3		5	
95	La sociedad líquida, de la que nos habla Bauman, parece enormemente expuesta. ~	REAL	1		3	
		PREDICHO	1		3	
96	un borrado de nuestros ahorros o de nuestras inversiones. ~	REAL	0		1	
		PREDICHO	0		1	
97	pero tal como van las cosas, mucho me temo que no.~	REAL	0		1	
		PREDICHO	0		1	1
98	guste a unos su formato o no guste. ~	REAL	0		2	
		PREDICHO	0		2	
99	eso hablando de instituciones públicas que, al menos, parecen cibernéticamente seguras. ~	REAL	0		1	
		PREDICHO	0		1	1
100	eso hablando de instituciones públicas que, al menos, parecen cibernéticamente seguras. ~	REAL	0		1	
		PREDICHO	0		1	1