
**IRIS: Asistente virtual para la redacción personalizada
de correos electrónicos**

IRIS: Virtual Assistant for Personalized Email Writing



CENTRO UNIVERSITARIO
DE TECNOLOGÍA Y ARTE DIGITAL

**Trabajo de Fin de Máster
Curso 2020–2021**

Autor
Carlos Moreno Morera

Director
Carlos Rodríguez Abellán

Máster en Data Science y Big Data
U-Tad

IRIS: Asistente virtual para la redacción personalizada de correos electrónicos

IRIS: Virtual Assistant for Personalized Email Writing

Trabajo de Fin de Máster en Data Science y Big Data

Autor
Carlos Moreno Morera

Director
Carlos Rodríguez Abellán

Convocatoria: Septiembre 2021
Calificación: *Nota*

Máster en Data Science y Big Data
U-Tad

26 de septiembre de 2021

Dedicatoria

*A Pedro Pablo y Marco Antonio, por crear TeXiS
e iluminar nuestro camino*

Agradecimientos

A Guillermo, por el tiempo empleado en hacer estas plantillas. A Adrián, Enrique y Nacho, por sus comentarios para mejorar lo que hicimos. Y a Narciso, a quien no le ha hecho falta el Anillo Único para coordinarnos a todos.

Resumen

IRIS: Asistente virtual para la redacción personalizada de correos electrónicos

Un resumen en castellano de media página, incluyendo el título en castellano. A continuación, se escribirá una lista de no más de 10 palabras clave.

Palabras clave

Máximo 10 palabras clave separadas por comas

Abstract

IRIS: Virtual Assistant for Personalized Email Writing

An abstract in English, half a page long, including the title in English. Below, a list with no more than 10 keywords.

Keywords

10 keywords max., separated by commas.

Índice

1. Introducción	1
1.1. Motivación	2
1.2. Objetivos	2
1.3. Plan de trabajo	2
1.4. Explicaciones adicionales sobre el uso de esta plantilla	2
1.4.1. Texto de prueba	2
2. Estado de la Cuestión	7
2.1. Correo electrónico	7
2.1.1. MIME	8
2.1.2. SMTP	12
2.1.3. POP	12
2.1.4. IMAP	13
2.2. Generación de Lenguaje Natural	14
2.2.1. ¿Qué es la Generación de Lenguaje Natural?	15
2.2.2. Arquitecturas para la Generación de Lenguaje Natural	16
3. Descripción del Trabajo	21
3.1. Enron corpus	22
4. Conclusiones y Trabajo Futuro	23
5. Introduction	25
6. Conclusions and Future Work	27
Bibliografía	29
A. Título del Apéndice A	37
B. Título del Apéndice B	39

Índice de figuras

2.1.	Estructura arbórea de tipos MIME de un e-mail de ejemplo	10
2.2.	Mensaje MIME de la figura 2.1	11
2.3.	Porcentaje de tiempo de un trabajador dedicado a cada tarea	14
2.4.	Arquitectura modular secuencial propuesta por Reiter y Dale (2000) para la NLG	19
2.5.	Ejemplo de mensaje	20
3.1.	Ejemplo de imagen	21

Índice de tablas

2.1. Previsión de usuarios de correo electrónico en todo el mundo (2021-2025)	8
2.2. Tráfico diario de correos electrónicos en todo el mundo (2021-2025)	8
3.1. Tabla de ejemplo	21

Capítulo 1

Introducción

“Frase célebre dicha por alguien inteligente”
— Autor

El estudiante elaborará una memoria descriptiva del trabajo realizado, con una **extensión mínima recomendada de 50 páginas** incluyendo al menos una introducción, objetivos y plan de trabajo, resultados con una discusión crítica y razonada de los mismos, conclusiones y bibliografía empleada en la elaboración de la memoria.

La memoria se puede redactar en castellano o en inglés, pero en el primer caso la introducción y las conclusiones de la memoria tienen que traducirse también al inglés y aparecerán como capítulos **al final de la memoria**. En ambos casos, el título de la memoria aparecerá en castellano y en inglés.

Además del cuerpo principal describiendo el trabajo realizado, la memoria contendrá los siguientes elementos, que no computarán para el cálculo de la extensión mínima del trabajo:

- un resumen en inglés de media página, incluyendo el título en inglés,
- ese mismo resumen en castellano, incluyendo el título en castellano,
- una lista de no más de 10 palabras clave en inglés,
- esa misma lista en castellano,
- un índice de contenidos, y
- una bibliografía.

La portada de la memoria deberá contener la siguiente información:

- "Máster en NOMBRE DEL MÁSTER, Facultad de Informática, Universidad Complutense de Madrid"
- Título
- Autor
- Director(es)
- Colaborador externo de dirección, si lo hay

- Curso académico
- Solo en la versión final: convocatoria y calificación obtenida

Para facilitar la escritura de la memoria siguiendo esta estructura, el estudiante podrá usar las plantillas en LaTeX o Word preparadas al efecto y publicadas en la página web del máster correspondiente.

Todo el material no original, ya sea texto o figuras, deberá ser convenientemente citado y referenciado. En el caso de material complementario se deben respetar las licencias y copyrights asociados al software y hardware que se emplee. En caso contrario no se autorizará la defensa, sin menoscabo de otras acciones que correspondan.

1.1. Motivación

Introducción al tema del TFM.

1.2. Objetivos

Descripción de los objetivos del trabajo.

1.3. Plan de trabajo

Aquí se describe el plan de trabajo a seguir para la consecución de los objetivos descritos en el apartado anterior.

1.4. Explicaciones adicionales sobre el uso de esta plantilla

Si quieras cambiar el **estilo del título** de los capítulos, edita `TeXiS\TeXiS_pream.tex` y comenta la línea `\usepackage[Lenny]{fncychap}` para dejar el estilo básico de `LATEX`.

Si no te gusta que no haya **espacios entre párrafos** y quieres dejar un pequeño espacio en blanco, no metas saltos de línea (`\\\`) al final de los párrafos. En su lugar, busca el comando `\setlength{\parskip}{0.2ex}` en `TeXiS\TeXiS_pream.tex` y aumenta el valor de `0,2ex` a, por ejemplo, `1ex`.

TFMTeXiS se ha elaborado a partir de la plantilla de TeXiS¹, creada por Marco Antonio y Pedro Pablo Gómez Martín para escribir su tesis doctoral. Para explicaciones más extensas y detalladas sobre cómo usar esta plantilla, recomendamos la lectura del documento `TeXiS-Manual-1.0.pdf` que acompaña a esta plantilla.

El siguiente texto se genera con el comando `\lipsum[2-20]` que viene a continuación en el fichero `.tex`. El único propósito es mostrar el aspecto de las páginas usando esta plantilla. Quita este comando y, si quieres, comenta o elimina el paquete `lipsum` al final de `TeXiS\TeXiS_pream.tex`

1.4.1. Texto de prueba

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet,

¹<http://gaia.fdi.ucm.es/research/texis/>

tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellenesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellenesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellenesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetur at, consectetur sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a,

dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetur a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetur. Nullam elementum, urna vel imperdierit sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdierit lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetur odio sem sed wisi.

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetur eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdierit justo nec dolor.

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetur tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdierit. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdierit sed, pellentesque nec, nisl. Vestibulum imperdierit neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero. Suspendisse in elit quis nisl aliquam dapibus. Pellentesque auctor sapien. Sed egestas sapien nec lectus. Pellentesque vel dui vel neque bibendum viverra. Aliquam porttitor nisl nec pede. Proin mattis libero vel turpis. Donec rutrum mauris et libero. Proin euismod porta felis. Nam lobortis, metus quis elementum commodo, nunc lectus elementum mauris, eget vulputate ligula tellus eu neque. Vivamus eu dolor.

Nulla in ipsum. Praesent eros nulla, congue vitae, euismod ut, commodo a, wisi. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Aenean nonummy magna non leo. Sed felis erat, ullamcorper in, dictum non, ultricies ut, lectus. Proin vel arcu a odio lobortis euismod. Vestibulum ante ipsum primis in faucibus

orci luctus et ultrices posuere cubilia Curae; Proin ut est. Aliquam odio. Pellentesque massa turpis, cursus eu, euismod nec, tempor congue, nulla. Duis viverra gravida mauris. Cras tincidunt. Curabitur eros ligula, varius ut, pulvinar in, cursus faucibus, augue.

Nulla mattis luctus nulla. Duis commodo velit at leo. Aliquam vulputate magna et leo. Nam vestibulum ullamcorper leo. Vestibulum condimentum rutrum mauris. Donec id mauris. Morbi molestie justo et pede. Vivamus eget turpis sed nisl cursus tempor. Curabitur mollis sapien condimentum nunc. In wisi nisl, malesuada at, dignissim sit amet, lobortis in, odio. Aenean consequat arcu a ante. Pellentesque porta elit sit amet orci. Etiam at turpis nec elit ultricies imperdiet. Nulla facilisi. In hac habitasse platea dictumst. Suspendisse viverra aliquam risus. Nullam pede justo, molestie nonummy, scelerisque eu, facilisis vel, arcu.

Curabitur tellus magna, porttitor a, commodo a, commodo in, tortor. Donec interdum. Praesent scelerisque. Maecenas posuere sodales odio. Vivamus metus lacus, varius quis, imperdiet quis, rhoncus a, turpis. Etiam ligula arcu, elementum a, venenatis quis, sollicitudin sed, metus. Donec nunc pede, tincidunt in, venenatis vitae, faucibus vel, nibh. Pellentesque wisi. Nullam malesuada. Morbi ut tellus ut pede tincidunt porta. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam congue neque id dolor.

Donec et nisl at wisi luctus bibendum. Nam interdum tellus ac libero. Sed sem justo, laoreet vitae, fringilla at, adipiscing ut, nibh. Maecenas non sem quis tortor eleifend fermentum. Etiam id tortor ac mauris porta vulputate. Integer porta neque vitae massa. Maecenas tempus libero a libero posuere dictum. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aenean quis mauris sed elit commodo placerat. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Vivamus rhoncus tincidunt libero. Etiam elementum pretium justo. Vivamus est. Morbi a tellus eget pede tristique commodo. Nulla nisl. Vestibulum sed nisl eu sapien cursus rutrum.

Nulla non mauris vitae wisi posuere convallis. Sed eu nulla nec eros scelerisque pharetra. Nullam varius. Etiam dignissim elementum metus. Vestibulum faucibus, metus sit amet mattis rhoncus, sapien dui laoreet odio, nec ultricies nibh augue a enim. Fusce in ligula. Quisque at magna et nulla commodo consequat. Proin accumsan imperdiet sem. Nunc porta. Donec feugiat mi at justo. Phasellus facilisis ipsum quis ante. In ac elit eget ipsum pharetra faucibus. Maecenas viverra nulla in massa.

Nulla ac nisl. Nullam urna nulla, ullamcorper in, interdum sit amet, gravida ut, risus. Aenean ac enim. In luctus. Phasellus eu quam vitae turpis viverra pellentesque. Duis feugiat felis ut enim. Phasellus pharetra, sem id porttitor sodales, magna nunc aliquet nibh, nec blandit nisl mauris at pede. Suspendisse risus risus, lobortis eget, semper at, imperdiet sit amet, quam. Quisque scelerisque dapibus nibh. Nam enim. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc ut metus. Ut metus justo, auctor at, ultrices eu, sagittis ut, purus. Aliquam aliquam.

Capítulo 2

Estado de la Cuestión

En el estado de la cuestión es donde aparecen gran parte de las referencias bibliográficas del trabajo. Una de las formas más cómodas de gestionar la bibliografía en L^AT_EX es utilizando **bibtex**. Las entradas bibliográficas deben estar en un fichero con extensión *.bib* (con esta plantilla se proporciona el fichero *biblio.bib*, donde están las entradas referenciadas más abajo).

2.1. Correo electrónico

El correo electrónico (Guide, 2005, Capítulo 11) es un servicio de comunicación que ha sido utilizado desde 1971 (Ibrahim et al., 2018), momento en el que a través de la primera red adaptada para el envío de e-mails se envió el texto “QWERTYUIOP”. Este correo se mandó a través de ARPAnet (cuyo nombre proviene de *Advanced Research Projects Agency Network*, que en inglés significa Red de la Agencia de Proyectos de Investigación Avanzada, y fue la primera red en la que se implementó el famoso protocolo TCP/IP) con un protocolo experimental conocido como CYPNET. Actualmente los mensajes hacen uso de una arquitectura cliente-servidor, de manera que el correo electrónico es construido a través de un programa cliente y, posteriormente, es enviado al servidor. Desde dicho servidor, se redirige el mensaje al servidor del servicio de correo del destinatario y, desde este último, es enviado al receptor.

De acuerdo con Radicati y Levenstein (2021), el correo electrónico “sigue siendo la forma de comunicación dominante tanto para las empresas como para los consumidores particulares” y, aún hoy en día, cada año se continúa observando un constante crecimiento del número de cuentas de e-mail y de la cantidad de mensajes enviados. De hecho, en 2021 el número de usuarios de correo electrónico en todo el mundo alcanzará los 4.1 miles de millones (más de la mitad de la población mundial utiliza el servicio de e-mail) y se espera que esta cifra siga aumentando hasta que haya 4.5 miles de millones en 2025. El crecimiento del número de usuarios en este rango de años se ve reflejado en la tabla 2.1.

Por otro lado, la evolución en el tráfico diario de e-mails en todo el mundo se presenta en la tabla 2.2, donde puede observarse la gigantesca cantidad de correos electrónicos enviados cada día y su crecimiento a lo largo de los próximos cuatro años. Con estos datos, podemos calcular el número medio de mensajes enviados por usuario cada día, obteniendo que en 2021 de media cada usuario manda aproximadamente 77 e-mails y esta cantidad continúa creciendo hasta alcanzar casi los 82 correos electrónicos diarios de media por usuario. Esto significa que, a medida que avanzan los años, no solo crece la cifra de personas que hace

Año	2021	2022	2023	2024	2025
Miles de millones de usuarios en todo el mundo	4.147	4.258	4.371	4.481	4.594
Porcentaje de crecimiento	3 %	3 %	3 %	3 %	3 %

Tabla 2.1: Previsión de usuarios de correo electrónico en todo el mundo (2021-2025)

Tabla extraída de Radicati y Levenstein (2021).

uso de este sistema de comunicación, sino que también aumenta la dedicación que cada usuario invierte en la utilización de esta herramienta.

Año	2021	2022	2023	2024	2025
Miles de millones de correos electrónicos enviados/recibidos al día en el mundo	319.6	333.2	347.3	361.6	376.4
Porcentaje de crecimiento	4.3 %	4.3 %	4.2 %	4.1 %	4.1 %

Tabla 2.2: Tráfico diario de correos electrónicos en todo el mundo (2021-2025)

Tabla extraída de Radicati y Levenstein (2021).

Para hacer posible el envío de todos estos correos electrónicos, existe un estándar que determina el formato que deben tener los mensajes y una amplia gama de protocolos de red que permiten el intercambio de e-mails entre máquinas distintas (las cuales a menudo poseen sistemas operativos distintos y utilizan diferentes programas de correo electrónico). A continuación se presenta dicho estándar de formato conocido como MIME (véase la sección 2.1.1), el cual resultará de gran utilidad de cara a procesar cada uno de los mensajes pertenecientes corpus inicial de partida (explicado en 3.1) y obtener la información necesaria de cada uno de ellos. También, con el fin de cerrar este apartado y tener un conocimiento general acerca del funcionamiento de este medio de comunicación, se introducirán los principales protocolos de gestión de correos electrónicos tanto para la transmisión de los mismos (para dicha tarea se hace uso del protocolo SMTP expuesto en la sección 2.1.2) como para el acceso por parte de los usuarios (en este caso se utilizan los protocolos POP e IMAP que son explicados en las secciones 2.1.3 y 2.1.4, respectivamente).

2.1.1. MIME

La especificación del formato que deben tener los correos electrónicos viene determinada por el estándar conocido como MIME (acrónimo de *Multipurpose Internet Mail Extensions*), el cual es utilizado para el intercambio de distintos tipos de archivos (texto, audio y vídeos, entre otros) que ofrece soporte a textos con caracteres no pertenecientes al formato ASCII, archivos adjuntos que no son de texto, mensajes con cuerpo con numerosas partes (conocidos como mensajes multiparte) e información de cabecera con caracteres no ASCII. Se encuentra definido en los documentos técnicos llamados *Request For Comments* (RFC) con identificadores: RFC 2045 (Freed y Borenstein, 1996b), RFC 2046 (Freed y Borenstein, 1996c), RFC 2047 (Moore, 1996), RFC 2049 (Freed y Borenstein, 1996a), RFC 2077 (Nelson y Parks, 1997), RFC 4288 (Freed y Klensin, 2005a) y RFC 4289 (Freed y Klensin, 2005b).

Prácticamente todos los correos electrónicos escritos por personas en Internet y una considerable proporción de estos mensajes generados automáticamente, se transmiten en formato MIME a través de SMTP (véase la sección 2.1.2). Los mensajes de correo electrónico de Internet están tan estrechamente relacionados con SMTP y MIME que suelen

denominarse mensajes SMTP/MIME.

Los tipos de contenido englobados dentro del estándar MIME son de gran importancia también fuera del contexto de los correos electrónicos. Ejemplos de ello son algunos protocolos de red como el HTTP de la Web. Este protocolo requiere que los datos se transmitan en un contexto de mensaje de tipo e-mail, aunque los datos no sean un correo electrónico propiamente dicho.

Hoy en día, ningún programa de correo electrónico o navegador de Internet puede considerarse completo si no acepta MIME en sus distintas funcionalidades (formatos de texto y de archivo).

2.1.1.1. Nomenclatura de tipos

Como se ha mencionado anteriormente, MIME permite el intercambio de distintos tipos de archivos. Para lograrlo, este estándar utiliza una nomenclatura diferente para denotar a cada tipo. Los nombres utilizados siguen el formato “tipo/subtipo”, siendo tanto tipo como subtipo cadenas de caracteres. De esta manera, el tipo especificará la categoría general de los datos enviados y el subtipo determinará el tipo específico de la información mandada. Los valores que puede tomar tipo son los siguientes:

- *text*: informa de que el contenido es texto. Este tipo puede preceder a los subtipos *html*, *xml* y *plain*.
- *multipart*: indica que el mensaje contiene distintas partes (cada una de un tipo diferente) con datos independientes entre ellas. Puede anteceder a subtipos como *form-data* y *digest*.
- *message*: se utiliza para encapsular un mensaje existente, por ejemplo, cuando se quiere responder a un correo electrónico y añadir los mensajes anteriores. A este tipo le pueden seguir subtipos como *partial* y *rfc822*.
- *image*: especifica que el contenido se trata de una imagen. Le pueden suceder los subtipos *png*, *jpeg* y *gif*.
- *audio*: determina que el contenido se trata de un audio. Los subtipos *mp3* y *32kadpcm* son algunos ejemplos a los que puede anteceder este tipo.
- *video*: señala que el contenido se trata de un vídeo. Puede preceder a subtipos como *mpeg* y *avi*.
- *application*: denota a los datos de aplicación que pueden ser binarios. Algunos de sus subtipos correspondientes son *json* y *pdf*.
- *font*: significa que el contenido del mensaje es un archivo que define el formato de una fuente. Le pueden suceder subtipos como *woff* y *ttf*.

2.1.1.2. Cabeceras MIME

Cuando se codifica un correo electrónico siguiendo el estándar MIME, se estructura en diferentes cabeceras cuyo valor asociado nos dará información acerca del mensaje enviado. Las cabeceras más importantes son:

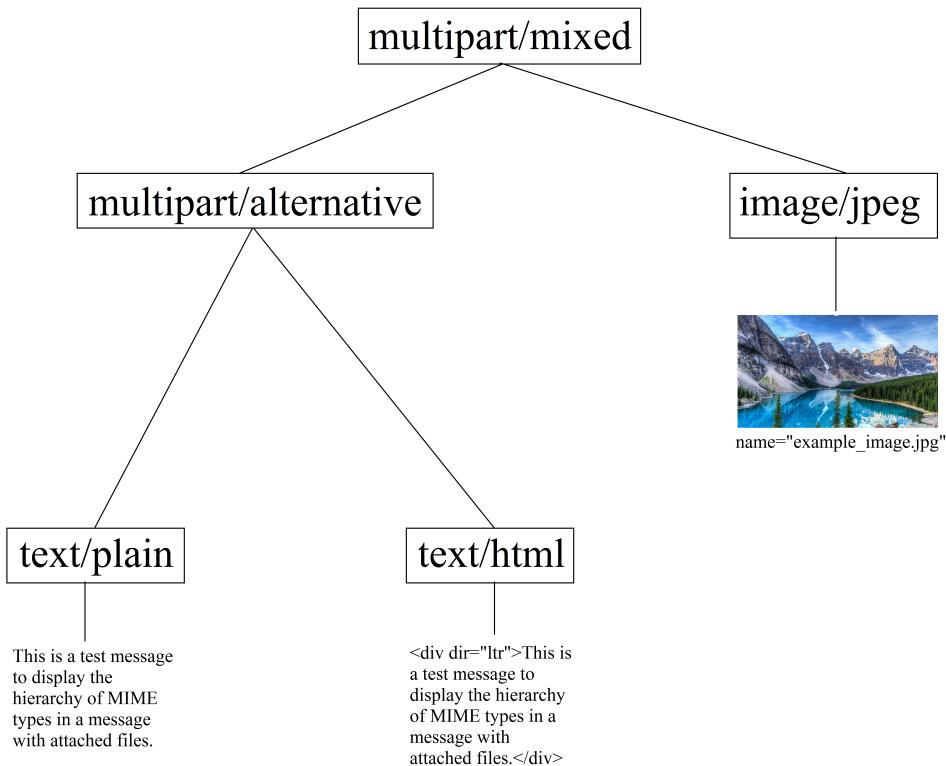


Figura 2.1: Estructura arbórea de tipos MIME de un e-mail de ejemplo
Imagen extraída de Moreno Morera (2020)

- *Content-Type*: el valor asociado a esta cabecera es el tipo y subtipo del mensaje con el mismo formato que se ha explicado anteriormente (véase la sección 2.1.1.1). Por ejemplo, si se observa la cabecera y el valor *Content-Type*: `text/plain`, indicará que el mensaje es un texto plano. El uso del tipo `multipart` hace posible la creación de correos electrónicos con partes y subpartes organizadas en una estructura arbórea, en la cual los nodos hoja pertenecen a cualquier tipo y el resto pueden tratarse de algún subtipo de `multipart` (Freed y Borenstein, 1992, Sección 7.2). Para poder entender mejor cómo se estructuran este tipo de e-mails, en la figura 2.1 se presenta un posible mensaje con una parte de texto plano, otra de texto HTML y una imagen. Para crear este correo electrónico, es necesario contar con un nodo raíz del tipo `multipart/mixed`. Además, como puede observarse en la figura 2.1, la utilización del tipo `multipart/alternative` permite la coexistencia en un mismo mensaje del cuerpo tanto en formato de texto plano como HTML. Sobra decir que, gracias a este formato de estructura en árbol de la cabecera *Content-Type*, es posible construir muchas otras variedades de mensajes (como adjuntar el mensaje original que ha sido reenviado utilizando `multipart/mixed` con una parte `text/plain` y otra `message/rfc822`).

Un detalle importante es el hecho de que si se comparan las figuras 2.1 y 2.2, se podrá observar que cada nodo de la estructura arbórea del correo electrónico (mostrada en la figura 2.1) es presentado en la figura 2.2 (que es como se reciben los e-mails realmente y como se encuentran almacenados en el corpus) habiendo recorrido el árbol en profundidad en preorden.

- *Content-Disposition*: esta cabecera indica la forma de presentación de la parte del

```

MIME-Version: 1.0
From: Sender <sender@gmail.com>
Date: Fri, 4 Oct 2019 22:06:29 +0200
Message-ID: <CADkvQ12BgJJxjqApD6AcNcAhJJEZ5FBkPi-9Pw_XduXgN_3sQ@mail.gmail.com>
Subject: Example MIME message
To: addressee@gmail.com
Cc: addressee2@gmail.com
Content-Type: multipart/mixed; boundary="000000000000abadb805941b3c9d"

--000000000000abadb805941b3c9d
Content-Type: multipart/alternative; boundary="000000000000abadb605941b3c9b"

--000000000000abadb605941b3c9b
Content-Type: text/plain; charset="UTF-8"
Content-Transfer-Encoding: quoted-printable

This is a test message to display the hierarchy of MIME types in a message
with attached files.

--000000000000abadb605941b3c9b
Content-Type: text/html; charset="UTF-8"
Content-Transfer-Encoding: quoted-printable

<div dir="ltr">This is a test message to display the hierarchy of MIME types in a message with attached
files.</div>

--000000000000abadb605941b3c9b--
--000000000000abadb805941b3c9d
Content-Type: image/jpeg; name="example_image.jpg"
Content-Disposition: attachment; filename="example_image.jpg"
Content-Transfer-Encoding: base64
Content-ID: <16d985155126855b6e01>
X-Attachment-Id: 16d985155126855b6e01

--000000000000abadb805941b3c9d--

```

Figura 2.2: Mensaje MIME de la figura 2.1
Imagen extraída de Moreno Morera (2020)

mensaje a la que pertenece. Puede tener dos posibles valores: *inline* (que se utiliza cuando el contenido debe ser mostrado al mismo tiempo que el cuerpo del mensaje, como por ejemplo cuando se inserta una imagen en el texto y no como archivo adjunto) y *attachment* (este valor determinará que la parte del mensaje requerirá algún tipo de acción por parte del usuario para visualizar el contenido, como por ejemplo en el caso de adjuntar un archivo). Además, esta cabecera dispone de distintos campos que reflejan más información acerca del contenido, como puede ser el nombre del archivo y la fecha de creación o de modificación. A continuación se presenta un ejemplo extraído de Troost et al. (1997) de esta cabecera y los campos que pueden acompañarle:

```

Content-Disposition: attachment; filename=genome.jpeg;
modification-date="Wed, 12 Feb 1997 16:29:51 -0500";

```

También puede observarse esta cabecera en la última parte del mensaje de ejemplo de la figura 2.2.

- *Content-Transfer-Encoding*: cuando se mandan archivos en un correo electrónico, a veces estos se codifican como 8-bit o archivos binarios, codificaciones no soportadas por determinados protocolos. Por este motivo, es necesario poseer un estándar que

especifique cómo debe re-codificarse este tipo de información en un formato 7-bit. La cabecera *Content-Transfer-Encoding* (Freed y Borenstein, 1992, Sección 5) indica al cliente qué tipo de transformación se ha efectuado para que este sea capaz de recuperar los datos originales. Los posibles valores son *base64* (Josefsson, 2006; Freed y Borenstein, 1996b), *quoted-printable* (Borenstein y Freed, 1993), *8bit*, *7bit*, *binary* y *x-token*. Todos ellos hacen referencia a un tipo de codificación que se encuentran fuera del alcance de este trabajo y para las que se recomienda consultar las referencias bibliográficas en caso de querer profundizar en ellas.

2.1.2. SMTP

El SMTP (cuyas siglas hacen referencia a *Simple Mail Transfer Protocol*) es un protocolo de red orientado a conexión utilizado para el intercambio de correos electrónicos. Originalmente fue definido por Postel (1982) (para especificar cómo llevar a cabo el envío de mensajes) y Crocker (1982) (que presenta el formato que deben tener los e-mails). Actualmente, se deben consultar los RFC desarrollados por Klensin (2008) y Resnick (2008) que, respectivamente, sustituyen a los dos originales.

Al ser un protocolo de transferencia de mensajes, posee algunas limitaciones a la hora de recibir e-mails en el servidor de destino. Por ello, esta tarea se delega a otros protocolos como el POP (véase la sección 2.1.3) e IMAP (véase la sección 2.1.4), mientras que el SMTP se encarga única y exclusivamente del envío.

Cuando se hace uso del SMTP un correo electrónico es enviado (esta acción de denota con la palabra *push*) de un servidor a otro hasta que alcanza su destino. El mensaje se encamina en función del servidor de correo de destino, en lugar de hacerlo en función de los destinatarios individuales del mensaje especificados durante la conexión del cliente al servidor SMTP. Gracias a que este protocolo dispone de una función para iniciar el procesamiento de la cola de correo, un servidor de correo conectado de forma intermitente puede extraer mensajes de otro servidor remoto cuando sea necesario.

2.1.3. POP

El POP (cuyas siglas hacen referencia a *Post Office Protocol*) es un protocolo de aplicación en el modelo OSI utilizado para la obtención de e-mails almacenados en un servidor remoto de Internet denominado servidor POP. Originalmente fue definido por Reynolds (1984), que especificó la primera versión de POP, también conocida como POP1. La versión actual de POP, POP3 (en general cuando se habla de POP se refiere a esta versión), fue detallada por Myers et al. (1996).

El protocolo POP posee numerosos comandos que hacen posible la conexión manual con el servidor POP3. Además, soporta otros como LIST, RETR y DELE, que permiten la gestión de los mensajes del usuario con acciones como mostrarlos, descargarlos o borrarlos, respectivamente.

POP3 fue diseñado para la tarea de recepción de correos electrónicos. Gracias a este protocolo, los usuarios con conexiones a Internet intermitentes o muy lentas pueden descargar sus mensajes mientras se encuentran conectados a la red y consultarlos estando *offline*. La sucesión de operaciones más común se produce cuando un cliente se conecta, descarga todos sus mensajes, los almacena en su dispositivo local como e-mails nuevos, se borran del servidor y, por último, el usuario se desconecta. Sin embargo, algunos clientes de correo incluyen la opción de dejar los mensajes almacenados en el servidor en lugar de borrarlos. Estos utilizan el comando UIDL (*Unique IDentification Listing*) el cual, a diferencia del resto de instrucciones de POP3, no identifica el correo electrónico a través del número

ordinal asociado por el servidor, ya que generaría problemas si un cliente tratara de dejar ciertos mensajes en el servidor debido a que este número cambiaría de una conexión a otra. En lugar de ello, asigna a cada mensaje un identificador constituido por una cadena de caracteres única y permanente. De esta manera, se puede determinar fácilmente qué mensajes se quieren almacenar en el servidor a la vez que se descargan.

Al igual que otros protocolos más antiguos, POP3 utiliza un mecanismo de firma sin encriptación. De hecho, la transmisión de contraseñas POP3 en texto plano continúa ocurriendo. Hoy en día, POP3 tiene varios métodos de autenticación que ofrecen un amplio rango de niveles de protección contra accesos ilegales a las bandejas de correo de los usuarios.

La ventaja de POP3 frente a otros protocolos es que entre el cliente y el servidor no es necesario mandar un gran número de comandos para comunicarse. Este protocolo también resulta muy útil cuando no se cuenta con una conexión constante a Internet o a la red que aloja al servidor.

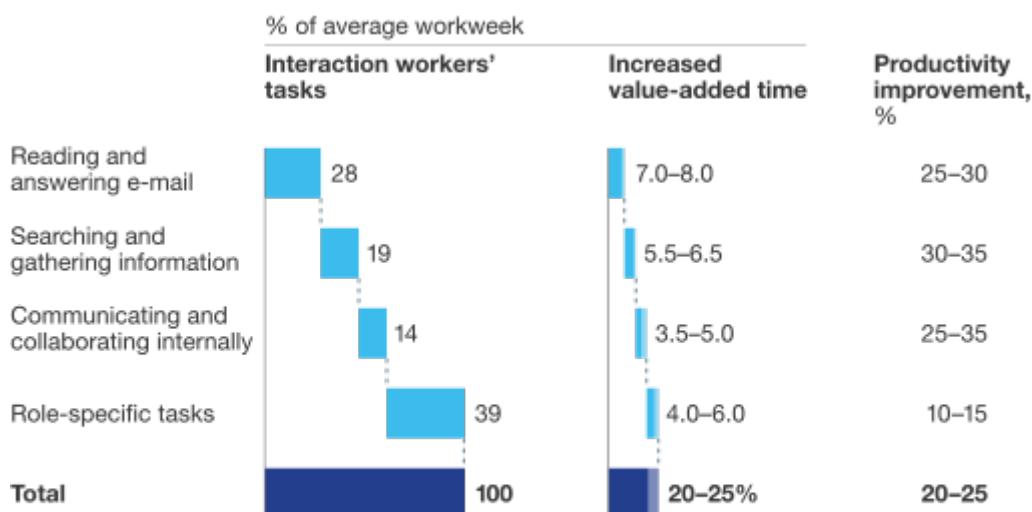
2.1.4. IMAP

El IMAP (cuyas siglas hacen referencia a *Internet Message Access Protocol*) es un protocolo de aplicación diseñado como alternativa al POP (véase la sección 2.1.3) en 1986, el cual permite el acceso a los mensajes almacenados en un servidor de Internet. Al igual que con el POP, con el IMAP es posible acceder a la cuenta de correo electrónico desde cualquier dispositivo con conexión a Internet. La versión actual del IMAP (IMAP versión 4 revisión 1 o IMAP4rev1) fue definida por Crispin (2003).

A diferencia del POP, el IMAP abre la puerta a la gestión de la misma bandeja de entrada por parte de múltiples clientes. Esta característica se produce gracias a las principales diferencias entre los dos protocolos: el IMAP no elimina los mensajes del servidor hasta que el cliente lo solicite explícitamente (mientras que el POP los borra por defecto, lo que hace imposible acceder a ellos desde otro dispositivo que no haya descargado previamente los correos electrónicos) y tampoco descarga los e-mails en el dispositivo del usuario, aunque opcionalmente es factible tener una copia local de los mismos. Esta última propiedad del IMAP da lugar a algunas ventajas frente al POP. Una de ellas es la posibilidad e notificar de manera inmediata de la llegada de un nuevo correo electrónico (ya que el IMAP funciona con una conexión cliente-servidor permanente), mientras que el POP verifica si hay nuevos mensajes cada pocos minutos (lo cual provoca un aumento apreciable del tráfico y del tiempo que el usuario tiene que esperar para enviar una solicitud al servidor, ya que es necesario completar primero la descarga de todos los mensajes nuevos). Por otro lado, gracias al IMAP, los usuarios pueden crear carpetas compartidas con otras personas (esta funcionalidad depende del servidor de correo) y los e-mails no ocupan espacio de memoria en el dispositivo local, mientras que el POP los descarga independientemente de si van a ser leídos o no (estrictamente hablando, el IMAP tiene que descargar el mensaje cuando va a ser leído, pero se trata de archivos temporales y solo se extraen las cabeceras de los correos electrónicos a la hora de gestionar la bandeja de entrada). Precisamente, el hecho de evitar la descarga del e-mail, permite al usuario gestostrar carpetas, plantillas y borradores en el servidor además de poder llevar a cabo una búsqueda en la bandeja de entrada mediante palabras clave.

2.2. Generación de Lenguaje Natural

Como se ha descrito y observado gracias a la tabla 2.2, el tráfico de correos electrónicos diarios continúa en constante crecimiento hasta llegar, al menos, a la gigantesca cifra de 376 mil millones enviados al día en todo el mundo. Esto se revierte en una gran cantidad de tiempo invertido para la redacción de todos estos mensajes que no se mandan de manera automática. Sin embargo, esta gran dedicación al e-mail lleva produciéndose desde hace más de una década, cuando no nos encontrábamos con cifras de tráfico tan elevadas. Según Chui et al. (2012), de media los empleados invertían el 28 % de su tiempo semanal en la gestión del correo electrónico (como viene reflejado en la figura 2.3). Esto se traduce en más de once horas dedicadas única y exclusivamente a leer y contestar mensajes, enviando y recibiendo una media de 124 e-mails por día (Radicati y Levenstein, 2015). Por si estos datos no fueran suficientemente preocupantes, de cara a la productividad laboral y resolución eficiente de las tareas, según Segal (2021) este problema se ha agravado en los últimos años por diversas causas (entre las que se encuentra la pandemia de la Covid-19). En definitiva, hoy en día podemos afirmar que tanto en el ámbito profesional como personal se invierte una gran cantidad de esfuerzo y tiempo para gestionar nuestra cuenta de e-mail, lo cual plantea un problema en el que vemos que, en lugar de ser una herramienta útil, se convierte en una responsabilidad más que debe llevarse al día y de la que no es posible desprenderte ya que la capacidad de mandar estos mensajes es imprescindible para llevar a cabo tareas del día a día. Pero, ¿y si fuera posible ahorrar todo este tiempo de escritura de correos electrónicos?



Source: International Data Corporation (IDC); McKinsey Global Institute analysis

Figura 2.3: Porcentaje de tiempo de un trabajador dedicado a cada tarea

Para lograr este propósito es imprescindible profundizar en la rama de la Inteligencia Artificial conocida como *Generación de Lenguaje Natural* (cuyas siglas son *NLG* por su nombre en inglés *Natural Language Generation*). Un buen ejemplo de aplicación de las técnicas de generación automática de textos son los 100.000 libros que Philip M. Parker puso a la venta en la plataforma *Amazon.com* incluyendo títulos de temáticas tan variadas como *El libro oficial del paciente sobre la estenosis espinal* (Parker, 2002), *Perspectivas mundiales de 2009 a 2014 de los envases de 60 miligramos de Fromage Frais* (Parker, 2008a), *Perspectivas de 2007 a 2012 de las tapetes de nudo, alfombras de baño y conjuntos*

que miden 6 pies por 9 pies o menos en la India (Parker, 2006) y *Tesauro Quechua - Inglés* (Parker, 2008b).

Resulta evidente que dicha cantidad de libros no pudieron ser escritos por Parker, sino que debió hacerse uso de técnicas de generación automática de textos. El algoritmo utilizado para dicho propósito, se engloba dentro de los métodos de generación conocidos como *text-to-text* (texto a texto en castellano), dado que este tipo de técnicas toman como entrada textos ya existentes (normalmente escritos a mano y no generados automáticamente) y producen un nuevo texto coherente como salida. Otras aplicaciones de este tipo de métodos son la traducción automática de un idioma a otro (Hutchins y Somers, 2009; Oettinger, 2013), el resumen automático de textos (Mani y Maybury, 2001; Nenkova y McKeown, 2011), la simplificación de textos complejos, ya sea para hacerlos más accesibles para un público de lectores de bajo nivel de alfabetización (Siddharthan, 2014; Bautista et al., 2011) o niños (Macdonald y Siddharthan, 2016), corrección automática de ortografía, gramática y texto (Kukich, 1992; Ng et al., 2014), generación automática de revisiones de artículos científicos (Bartoli et al., 2016), generación de paráfrasis dada una frase de entrada (Bannard y Callison-Burch, 2005), generación automática de preguntas con fines didácticos y educativos (Brown et al., 2005), generación automática de relatos dada una descripción conceptual de la historia deseada (Gervás et al., 2004) o reescritura de textos (en concreto correos electrónicos) con estilo en función del destinatario (Moreno Morera, 2020).

Además de estos métodos *text-to-text*, existen los llamados *data-to-text* (datos a texto), en los cuales en lugar de recibir un texto como entrada, se genera el lenguaje a partir de datos. Estos pueden ser de todo tipo para dar lugar a informes o resúmenes como pueden ser de índole climatológica (Goldberg et al., 1994a; Ramos-Soto et al., 2014), financiera (Plachouras et al., 2016), ingenieril, como por ejemplo el trabajo desarrollado por Yu et al. (2007) para generar resúmenes de datos recopilados por sensores en turbinas de gas, sanitaria (Hüske-Kraus, 2003; Banaee et al., 2013), como la investigación llevada a cabo por Portet et al. (2009) para obtener informes textuales a partir de datos de cuidados intensivos neonatales, o, incluso, deportivos (Theune et al., 2001; Chen y Mooney, 2008). Además de informes o resúmenes, también se utilizan los métodos *data-to-text* para otros propósitos como la composición de discursos narrativos para relatos de varios personajes a partir de partidas de ajedrez (Gervás, 2014), redacción de periódicos electrónicos a partir de datos de sensores (Molina et al., 2011), generación de texto que aborda problemas medioambientales como el seguimiento de la fauna (Siddharthan et al., 2012; Ponnampерuma et al., 2013), la información medioambiental personalizada (Wanner et al., 2015) y la mejora del compromiso de los ciudadanos científicos a través de los comentarios generados (Van der Wal et al., 2016) o producción de información interactiva sobre artefactos culturales (Stock et al., 2007), entre otros.

Debido a que el objetivo de este trabajo se centra en la generación de correos electrónicos a partir del asunto, exploraremos en detalle las técnicas de Generación de Lenguaje Natural y, en especial, los métodos *text-to-text*. Para profundizar en los algoritmos y arquitecturas empleados ante los problemas de tipo *data-to-text*, conviene consultar la investigación llevada a cabo por Gatt y Krahmer (2018), en la cual muestran el estado del arte de los trabajos realizados en este ámbito.

2.2.1. ¿Qué es la Generación de Lenguaje Natural?

Dado que tanto los sistemas *text-to-text* como *data-to-text* y todas sus aplicaciones mencionadas anteriormente pertenecen a la rama de Generación de Lenguaje Natural, esta

no debe definirse en función de la entrada del sistema, sino en la salida. Según Reiter y Dale (2000) la NLG es la conceptualización del “campo de la inteligencia artificial y la lingüística computacional que se centra en los sistemas informáticos que son capaces de producir textos comprensibles en inglés u otra lengua humana. [...] Como área de investigación, la NLG presenta una perspectiva única ante problemas fundamentales de la inteligencia artificial, la ciencia cognitiva y la interacción. Estos incluyen cuestiones como por ejemplo cómo deben ser representados y cómo debe razonarse con la lingüística y el dominio del conocimiento, qué significa que un texto esté correctamente redactado y cómo es la mejor forma de comunicar información entre las computadoras y los usuarios.” Por lo tanto, la Generación de Lenguaje Natural se puede definir como el ámbito que engloba el estudio de la producción de lenguaje no artificial, así como el diseño e implementación de algoritmos y sistemas computacionales cuyo resultado debe ser un texto que imite la forma en que los humanos se comunican verbalmente (Vicente et al., 2015), ya sea oralmente o por escrito (del Socorro Bernardos, 2007). Es decir, independientemente de la entrada recibida, se precisa el significado de NLG a partir de la salida esperada por el problema planteado. Tanto es así, que, como hemos visto, la entrada del sistema puede variar excesivamente (McDonald, 1993): desde textos (que son precisamente los sistemas text-to-text) hasta datos de todo tipo como partidas de ajedrez (Gervás, 2014), pictogramas (González Álvarez y López Pulido, 2019) e, incluso, vídeos (Thomason et al., 2014). Sin embargo, autores como Dušek et al. (2020) acotan la definición de los sistemas de NLG estableciendo que la entrada deben ser representaciones semánticas, obviando así la primera tarea de la arquitectura propuesta por Reiter y Dale (2000) conocida como macro planificación o determinación del contenido (se explicará en la sección 2.2.2.1), que es precisamente el punto en el que se generan dichas representaciones semánticas.

Cabe destacar que, aunque desde un principio hayamos diferenciado entre métodos text-to-text y data-to-text, ni los límites entre las dos aproximaciones ni la pertenencia de algunas técnicas a ellas se encuentran claramente definidos. Un ejemplo de ello podemos encontrarlo en la generación automática de resúmenes de textos. En principio se caracterizaría claramente como un sistema text-to-text. No obstante, al hacer frente a este problema se han desarrollado soluciones con las conocidas técnicas abstractivas (Genest y Lapalme, 2011), que, como explican Hahn y Mani (2000), a diferencia de los métodos de extracción evitan recoger las frases completas y se limitan a tomar unidades semánticas. Este tipo de técnicas usadas, por ejemplo, en la obtención de opiniones de reseñas para la posterior generación de frases nuevas (Labbé y Portet, 2012), también provienen de problemas data-to-text. A la inversa, un sistema data-to-text puede hacer uso de técnicas que principalmente son utilizadas en los casos de uso text-to-text (McIntyre y Lapata, 2009; Kondadadi et al., 2013). Por otro lado, podría parecer que los métodos de *deep learning* (Goodfellow et al., 2016) deben ser mayoritariamente utilizados en los problemas data-to-text utilizando el trabajo llevado a cabo por Mikolov et al. (2013). Sin embargo, se han desarrollado extensamente esta clase de soluciones para la NLG con gran variedad de arquitecturas como las redes neuronales recurrentes (Cho et al., 2014; Tang et al., 2016), muy a menudo combinadas con la memoria a corto plazo o LSTM (Chen et al., 2016), o las arquitecturas conocidas como *transformers* (Vaswani et al., 2017).

2.2.2. Arquitecturas para la Generación de Lenguaje Natural

Ante el problema de generación de lenguaje natural, se han propuesta diversas soluciones para abordarlo. Sin embargo, actualmente preponderan dos tipos de arquitecturas: la propuesta por Reiter y Dale (2000), también conocida como arquitectura *realizer* (to-

ma el nombre de una de sus fases), que divide la generación en distintas subtareas y las aborda por separado, y la presentada por Vaswani et al. (2017), también conocida como arquitectura *transformer*, que expone una arquitectura constituida mayoritariamente por redes neuronales. Aunque en este trabajo hayamos hecho uso de esta última debido a la dificultad de establecer un dominio de lenguaje en los correos electrónicos, a continuación se hará una breve introducción a ambas.

2.2.2.1. Arquitectura realizer

A pesar de que en el trabajo de Reiter y Dale (2000) se centren principalmente en los sistemas de generación de lenguaje natural de tipo *data-to-text*, esta arquitectura también ha sido utilizada por sistemas *text-to-text* como para la generación automática de resúmenes a través de métodos abstractivos (Genest y Lapalme, 2011). De hecho, incluso es posible hacer uso de la conceptualización de la entrada del sistema planteada por Reiter y Dale (2000) para este tipo de problemas.

La filosofía de esta arquitectura se centra en la modularización de las diferentes tareas a abordar durante la generación de lenguaje natural. De esta manera, cada módulo se enfrenta a un reto específico con el que debe lidiar y se conecta con el módulo anterior haciendo coincidir la salida del previo con la entrada del actual y, lo mismo, con el módulo posterior. Asimismo, se construye un *pipeline* o arquitectura en secuencia de tareas con cada uno de los módulos.

La entrada de esta arquitectura viene determinada por una tupla de cuatro elementos (k, c, u, d) donde cada uno de ellos puede representarse de distintas maneras. El primer elemento de la tupla es la *base de conocimiento* (la letra viene dada por su denominación ingles *knowledge source*). Se trata de la información acerca del dominio de nuestro sistema de generación de lenguaje natural, la cual suele consistir en un conjunto de bases de datos y bases de conocimiento, como las ontologías (Fensel, 2001), que nuestra aplicación puede consultar durante su ejecución. Tanto la representación como el contenido de la base de conocimiento son altamente dependientes del tipo de aplicación que queramos construir, por ejemplo, en el trabajo de Reiter et al. (2005) la base de conocimiento consiste en parámetros meteorológicos numéricos de un modelo de predicción NWP, mientras que en el desarrollo presentado por Reiter et al. (1995) se utiliza como base de conocimiento un sistema de representación de conocimiento del mismo tipo que KL-ONE (Brachman y Schmolze, 1989) estructurado de manera jerárquica mediante relaciones *is-a* y *part-of*, del cual se pueden extraer diversas propiedades de las entidades que lo componen. Precisamente esta gran variabilidad en tan solo la primera componente de la entrada de la arquitectura realizer, es lo que sustenta la afirmación de Reiter y Dale (2000) de que no es posible proporcionar una caracterización formal genérica de lo que es una base de conocimiento y dificulta el establecimiento de esta componente en el sistema desarrollado si se hubiera elegido esta opción de arquitectura, ya que los correos electrónicos versan de una amplia variedad de temáticas muy distintas.

La segunda componente de la tupla de entrada es el objetivo de la comunicación (en inglés *communicative goal*). Este describe el propósito del texto para el que se quiere generar. Es importante no confundirlo con el propósito general del sistema de generación de lenguaje natural. Por ejemplo, el objetivo final de sistema implementado por Turner et al. (2007) es generar resúmenes textuales de datos de predicción meteorológica numérica espacio-temporal. Sin embargo, su propósito comunicativo de una ejecución determinada es el de presentar predicciones meteorológicas de una localización geográfica y un momento temporal dados.

La letra u de la tupla se corresponde con el modelo de usuario (en inglés *user model*), el cual consiste en una caracterización del receptor o público objetivo al que va dirigido el texto generado. Al igual que con el propósito comunicativo, no debe confundirse con los espectadores del sistema. Por ejemplo, el sistema implementado por Reiter et al. (1999) tiene como objetivo dirigirse a personas que consumen tabaco, ya que trata de generar cartas que convengan a los pacientes con este hábito para que intenten superar su adicción. No obstante, no resulta adecuado utilizar el mismo tipo de técnicas de convicción y el mismo lenguaje para una persona que ha comenzado a fumar desde hace poco que a otra que lleva muchos años consumiendo tabaco. También, podría resultar interesante definir perfiles en función de la edad y otras características del paciente que permitirían personalizar aún más estas cartas. Son precisamente este tipo de propiedades las que englobaría el modelo de usuario que el sistema toma como entrada. No obstante, a pesar de que no existen demasiados ejemplos de trabajos que incluyan esta variable de entrada como lo desarrollan en su estudio Goldberg et al. (1994b), por la dificultad de variación de los textos en función de dichos perfiles, este problema se suele tratar de abordar a través del estudio de la estilometría (Moreno Morera, 2020).

El último componente de la entrada es la historia del discurso (en inglés *discourse history*), la cual consiste en un modelo de la información transmitida y los temas tratados en el texto producido hasta el momento de la ejecución del sistema. Esto permite a la aplicación conocer las entidades y propiedades ya mencionadas gracias a las cuales es posible hacer un uso adecuado de los recursos anafóricos como los pronombres. En los sistemas de generación de lenguaje natural de interacción única, es decir, aquellos cuya ejecución produce un único texto con independencia de los generados en ejecuciones previas, la historia discursiva comienza como una estructura de datos vacía y se construye y utiliza durante la redacción del texto a generar. Todo lo contrario son los sistemas de diálogo, como los *chatbots*, en los cuales la historia del discurso suele hacer referencia al registro de diálogo, utilizado como repositorio de información sobre las interacciones previas entre el usuario y la aplicación de NLG. En trabajos como el desarrollado por Milosavljevic et al. (1996), en el cual se espera que los usuarios interactúen con una serie de textos relacionados, la historia discursiva resulta ser una mezcla de la utilizada en los sistemas de interacción única con la preponderante en los sistemas de diálogo. De esta manera, esta última componente de la tupla de entrada facilita el hacer referencias en el nuevo texto a entidades o conceptos mencionados en el texto actual o previos, o utilizar marcadores discursivos como “como se ha mencionado anteriormente” en momentos en que el texto generado repite información ya presentada en los anteriores.

Aunque puede resultar obvio, es importante tener en cuenta que, en la mayoría de los casos, será necesario un preprocessado del texto o los datos de entrada (sobre todo si han de ser analizados e interpretados) para facilitar el trabajo a la arquitectura presentada por Reiter y Dale (2000), de manera que sea más sencillo utilizar la entrada en cada una de las fases (Han et al., 2011). Este primer preprocessado es plenamente opcional y depende del origen de la entrada que se vaya a utilizar.

Una vez se conoce la entrada y salida establecida para un sistema de generación de lenguaje natural, se puede comenzar la presentación de la arquitectura modular secuencial. Según Reiter y Dale (2000), el proceso de generación puede descomponerse en tres fases: la macro planificación, la micro planificación y la realización (que da nombre a la arquitectura por ser la última fase). Los módulos que implementan cada una de ellas se conectan entre sí como muestra la figura 2.4.

En términos generales, el trabajo del módulo de macro planificación es producir una especificación del contenido del texto y su estructura, mediante el uso del dominio y el co-



Figura 2.4: Arquitectura modular secuencial propuesta por Reiter y Dale (2000) para la NLG

nocimiento de la aplicación sobre qué información es la más adecuada teniendo en cuenta la tupla de entrada. Esta fase normalmente también requiere conocer cómo suelen estructurarse los documentos del dominio de nuestro sistema. Muchas de las técnicas empleadas para la implementación de la macro planificación suelen asemejarse a las utilizadas en el ámbito de los sistemas expertos.

Para que el texto sea coherente, es preciso estructurar el contenido del mismo en el orden correcto. Por este motivo, la salida del módulo de macro planificación, el plan del documento, suele implementarse como una estructura de datos, generalmente arbórea, donde en cada nodo se encapsula la información más importante que debe formar parte de un párrafo o frase, además de información de cómo se relaciona con el resto de nodos. Con el objetivo de generar esta salida, se dividen las tareas de este nodo en determinación del contenido y estructuración del documento.

La determinación del contenido es la primera tarea de la macro planificación y, por tanto, del proceso de generación. En ella el sistema decide qué información es relevante para ser incluida en el texto y cuál no. Por lo general, en los sistemas data-to-text, se puede extraer más información en los datos de la que se quiere transmitir y, por ese motivo, cobra importancia la capacidad de selección y existen numerosas investigaciones al respecto, como la de Yu et al. (2007). Aunque la determinación del contenido está presente en la mayoría de los sistemas de generación de lenguaje natural (Mellish et al., 2006), los enfoques suelen estar estrechamente relacionados con el dominio, de forma que se construye una estructura de datos ad hoc, denominada “mensajes”, con el fin de especificar la salida de esta tarea (un ejemplo conceptual de un mensaje para un sistema que genera textos sobre la situación meteorológica puede observarse en la figura 2.5). Este planteamiento de la determinación del contenido, dificulta el desarrollo de sistemas en los que el dominio es extenso, como es el caso de los correos electrónicos, ya que no es posible implementar una clase que englobe todas las casuísticas (como se verá en la sección [poner n^o](#), para resolver este problema se utilizará el concepto de *Information Items* el cual pertenece al ámbito del resumen automático de textos). Sin embargo, sí se ha tratado de encontrar otro tipo de definición a la salida de esta tarea, como ocurre en el trabajo de Guhe (2007), el cual presenta una explicación cognitivamente plausible e incremental de la determinación del contenido, basada en estudios sobre las descripciones de observadores de eventos dinámicos a medida que se desarrollan.

La elección del contenido que debe ser expresado en el texto depende de varios factores, entre los que se encuentran: el propósito comunicativo, el modelo de usuario, las restricciones de la salida por limitaciones del dominio (como la longitud máxima que debe tener un texto) y la fuente de información (o base de conocimiento) subyacente disponible (es importante, aunque parezca evidente, no pretender generar información sobre la que el sistema no posee conocimiento o no es capaz de deducirlo con sus módulos de razonamiento correspondientes).

```
[type: MonthlyRainfallMsg
  period: [month: 05
            year: 1996]
  rainfall: [type: RelativeVariation
              magnitude: [unit: millimetres
                           number: 2]
              direction: -]]]
```

Figura 2.5: Ejemplo de mensaje
Imagen extraída de Reiter y Dale (2000)

Capítulo 3

Descripción del Trabajo

Aquí comienza la descripción del trabajo realizado. Se deben incluir tantos capítulos como sea necesario para describir de la manera más completa posible el trabajo que se ha llevado a cabo. Como muestra la figura 3.1, está todo por hacer.



Figura 3.1: Ejemplo de imagen

Si te sirve de utilidad, puedes incluir tablas para mostrar resultados, tal como se ve en la tabla 3.1.

Col 1	Col 2	Col 3
3	3.01	3.50
6	2.12	4.40
1	3.79	5.00
2	4.88	5.30
4	3.50	2.90
5	7.40	4.70

Tabla 3.1: Tabla de ejemplo

3.1. Enron corpus

Para llevar a cabo este trabajo, se ha elegido el corpus conocido como Enron¹, dado que los correos electrónicos que contiene pertenecieron a la empresa con el mismo nombre. Precisamente se hicieron públicos tras una investigación legal llevada a cabo a esta compañía por parte de la Comisión Federal de Regulación de la Energía² de Estados Unidos.

Enron corpus contiene 517.401 correos electrónicos de 150 usuarios distintos. Además de la ventaja de la gran cantidad de elementos pertenecientes a este dataset, también ha sido elegido por encontrar diversos trabajos sobre este mismo conjunto de e-mails, como el llevado a cabo por Klimt y Yang (2004).

¹<http://www-2.cs.cmu.edu/~enron/>

²<https://www.ferc.gov/>

Capítulo 4

Conclusiones y Trabajo Futuro

Conclusiones del trabajo y líneas de trabajo futuro.

Antes de la entrega de actas de cada convocatoria, en el plazo que se indica en el calendario de los trabajos de fin de máster, el estudiante entregará en el Campus Virtual la versión final de la memoria en PDF. En la portada de la misma deberán figurar, como se ha señalado anteriormente, la convocatoria y la calificación obtenida. Asimismo, el estudiante también entregará todo el material que tenga concedido en préstamo a lo largo del curso.

Chapter 5

Introduction

Introduction to the subject area. This chapter contains the translation of Chapter 1.

Chapter 6

Conclusions and Future Work

Conclusions and future lines of work. This chapter contains the translation of Chapter 4.

Bibliografía

Y así, del mucho leer y del poco dormir, se le secó el cerebro de manera que vino a perder el juicio.

Miguel de Cervantes Saavedra

BANAEH, H., AHMED, M. U. y LOUTFI, A. Towards nlg for physiological data monitoring with body area networks. En *14th European Workshop on Natural Language Generation, Sofia, Bulgaria, August 8-9, 2013*, páginas 193–197. 2013.

BANNARD, C. y CALLISON-BURCH, C. Paraphrasing with bilingual parallel corpora. En *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, páginas 597–604. 2005.

BARTOLI, A., DE LORENZO, A., MEDVET, E. y TARLAO, F. Your paper has been accepted, rejected, or whatever: Automatic generation of scientific paper reviews. En *International Conference on Availability, Reliability, and Security*, páginas 19–28. Springer, 2016.

BAUTISTA, S., LEÓN, C., HERVÁS, R. y GERVÁS, P. Empirical identification of text simplification strategies for reading-impaired people. En *Everyday Technology for Independence and Care*, páginas 567–574. IOS Press, 2011.

BORENSTEIN, N. y FREED, N. Mime (multipurpose internet mail extensions) part one: Mechanisms for specifying and describing the format of internet message bodies. Informe Técnico RFC 1521, Internet Engineering Task Force (IETF), 1993.

BRACHMAN, R. J. y SCHMOLZE, J. G. An overview of the kl-one knowledge representation system. *Readings in artificial intelligence and databases*, páginas 207–230, 1989.

BROWN, J., FRISHKOFF, G. y ESKENAZI, M. Automatic question generation for vocabulary assessment. En *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, páginas 819–826. 2005.

CHEN, D. L. y MOONEY, R. J. Learning to sportscast: a test of grounded language acquisition. En *Proceedings of the 25th international conference on Machine learning*, páginas 128–135. 2008.

- CHEN, Q., ZHU, X., LING, Z., WEI, S., JIANG, H. y INKPEN, D. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*, 2016.
- CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H. y BENGIO, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- CHUI, M., MANYIKA, J., BUGHIN, J., DOBBS, R., ROXBURGH, C., SARRAZIN, H., SANDS, G. y WESTERGREN, M. The social economy: Unlocking value and productivity through social technologies. *McKinsey Global Institute*, 2012.
- CRISPIN, M. Internet message access protocol - version 4rev1. Informe Técnico RFC 3501, University of Washington, 2003.
- CROCKER, D. H. Standard for the format of arpa internet text messages. Informe Técnico RFC 822, Dept. of Electrical Engineering, University of Delaware, 1982.
- DUŠEK, O., NOVIKOVA, J. y RIESER, V. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language*, vol. 59, páginas 123–156, 2020.
- FENSEL, D. Ontologies. En *Ontologies*, páginas 11–18. Springer, 2001.
- FREED, N. y BORENSTEIN, N. Mime (multipurpose internet mail extensions). Informe Técnico RFC 1341, Internet Engineering Task Force (IETF), 1992.
- FREED, N. y BORENSTEIN, N. Multipurpose internet mail extensions (mime) part five: Conformance criteria and examples. Informe Técnico RFC 2049, Internet Engineering Task Force (IETF), 1996a.
- FREED, N. y BORENSTEIN, N. Multipurpose internet mail extensions (mime) part one: Format of internet message bodies. Informe Técnico RFC 2045, Internet Engineering Task Force (IETF), 1996b.
- FREED, N. y BORENSTEIN, N. Multipurpose internet mail extensions (mime) part two: Media types. Informe Técnico RFC 2046, Internet Engineering Task Force (IETF), 1996c.
- FREED, N. y KLENSIN, J. Media type specifications and registration procedures. Informe Técnico RFC 4288, Internet Engineering Task Force (IETF), 2005a.
- FREED, N. y KLENSIN, J. Multipurpose internet mail extensions (mime) part four: Registration procedures. Informe Técnico RFC 4289, Internet Engineering Task Force (IETF), 2005b.
- GATT, A. y KRAHMER, E. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, vol. 61, páginas 65–170, 2018.
- GENEST, P.-E. y LAPALME, G. Framework for abstractive summarization using text-to-text generation. En *Proceedings of the workshop on monolingual text-to-text generation*, páginas 64–73. 2011.
- GERVÁS, P. Composing narrative discourse for stories of many characters: a case study over a chess game. *Literary and Linguistic Computing*, vol. 29(4), páginas 511–531, 2014.

- GERVÁS, P., DÍAZ-AGUDO, B., PEINADO, F. y HERVÁS, R. Story plot generation based on cbr. En *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, páginas 33–46. Springer, 2004.
- GOLDBERG, E., DRIEDGER, N. y KITTREDGE, R. I. Using natural-language processing to produce weather forecasts. *IEEE Expert*, vol. 9(2), páginas 45–53, 1994a.
- GOLDBERG, E., KITTREDGE, R. y DRIEDGER, N. Fog: a new approach to the synthesis of weather forecast text. *IEEE Expert (Special Track on NLP)*, 1994b.
- GONZÁLEZ ÁLVAREZ, S. y LÓPEZ PULIDO, J. M. Traductor de pictogramas a texto. 2019.
- GOODFELLOW, I., BENGIO, Y. y COURVILLE, A. *Deep learning*. MIT press, 2016.
- GUHE, M. Incremental conceptualization for language production. 2007.
- GUIDE, S. *Red Hat Enterprise Linux 4: Reference Guide*. Red Hat Inc., 2005. <http://web.mit.edu/rhel-doc/OldFiles/4/RH-DOCS/rhel-rg-en-4/index.html>.
- HAHN, U. y MANI, I. The challenges of automatic summarization. *Computer*, vol. 33(11), páginas 29–36, 2000.
- HAN, J., PEI, J. y KAMBER, M. *Data mining: concepts and techniques*. Elsevier, 2011.
- HÜSKE-KRAUS, D. Text generation in clinical medicine—a review. *Methods of information in medicine*, vol. 42(01), páginas 51–60, 2003.
- HUTCHINS, W. J. y SOMERS, H. L. *An introduction to machine translation*. 2009.
- IBRAHIM, M. S., KASIM, S., HASSAN, R., MAHDIN, H., RAMLI, A. A., FUDZEE, M. F. M., SALAMAT, M. A. ET AL. Information technology club management system. *Acta Electronica Malaysia*, vol. 2(2), páginas 01–05, 2018.
- JOSEFSSON, S. The base16, base32, and base64 data encodings. Informe Técnico RFC 4648, Internet Engineering Task Force (IETF), 2006.
- KLENSIN, J. Simple mail transfer protocol. Informe Técnico RFC 5321, Internet Engineering Task Force (IETF), 2008.
- KLIMT, B. y YANG, Y. Introducing the enron corpus. En *CEAS*. 2004.
- KONDADADI, R., HOWALD, B. y SCHILDER, F. A statistical nlg framework for aggregated planning and realization. En *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 1406–1415. 2013.
- KUKICH, K. Techniques for automatically correcting words in text. *Acm Computing Surveys (CSUR)*, vol. 24(4), páginas 377–439, 1992.
- LABBÉ, C. y PORTET, F. Towards an abstractive opinion summarisation of multiple reviews in the tourism domain. En *The First International Workshop on Sentiment Discovery from Affective Data (SDAD 2012)*, páginas 87–94. 2012.
- MACDONALD, I. y SIDDHARTHAN, A. Summarising news stories for children. ACL, 2016.
- MANI, I. y MAYBURY, M. T. Automatic summarization. 2001.

- MCDONALD, D. D. Issues in the choice of a source for natural language generation. *Computational Linguistics*, vol. 19(1), páginas 191–197, 1993.
- MCINTYRE, N. y LAPATA, M. Learning to tell tales: A data-driven approach to story generation. En *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, páginas 217–225. 2009.
- MELLISH, C., SCOTT, D., CAHILL, L., PAIVA, D., EVANS, R. y REAPE, M. A reference architecture for natural language generation systems. *Natural language engineering*, vol. 12(1), páginas 1–34, 2006.
- MIKOLOV, T., CHEN, K., CORRADO, G. y DEAN, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- MILOSAVLJEVIC, M., TULLOCH, A. y DALE, R. Text generation in a dynamic hypertext environment. *Australian Computer Science Communications*, vol. 18, páginas 417–426, 1996.
- MOLINA, M., STENT, A. y PARODI, E. Generating automated news to explain the meaning of sensor data. En *International Symposium on Intelligent Data Analysis*, páginas 282–293. Springer, 2011.
- MOORE, K. Multipurpose internet mail extensions (mime) part three: Message header extensions for non-ascii text. Informe Técnico RFC 2047, Internet Engineering Task Force (IETF), 1996.
- MORENO MORERA, C. Un modelo de análisis estilométrico de correos electrónicos para la redacción personalizada basada en el destinatario. 2020.
- MYERS, J., MELLON, C. y ROSE, M. Post office protocol - version 3. Informe Técnico RFC 1939, Dover Beach Consulting, Inc., 1996.
- NELSON, S. y PARKS, C. The model primary content type for multipurpose internet mail extensions. Informe Técnico RFC 2077, Internet Engineering Task Force (IETF), 1997.
- NENKOVA, A. y MCKEOWN, K. *Automatic summarization*. Now Publishers Inc, 2011.
- NG, H. T., WU, S. M., BRISCOE, T., HADIWINOTO, C., SUSANTO, R. H. y BRYANT, C. The conll-2014 shared task on grammatical error correction. En *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, páginas 1–14. 2014.
- OETTINGER, A. G. *Automatic language translation*. Harvard University Press, 2013.
- PARKER, P. M. *The Official Patient's Sourcebook on Spinal Stenosis*. Icon Group International Incorporated, 2002.
- PARKER, P. M. *The 2007-2012 Outlook for Tufted Washable Scatter Rugs, Bathmats, and Sets That Measure 6-Feet by 9-Feet or Smaller in India*. Icon Group International Incorporated, 2006.
- PARKER, P. M. *The 2009-2014 World Outlook for 60-milligram Containers of Fromage Frais*. Icon Group International Incorporated, 2008a.

- PARKER, P. M. *Webster's Quechua - English Thesaurus Dictionary*. Icon Group International Incorporated, 2008b.
- PLACHOURAS, V., SMILEY, C., BRETZ, H., TAYLOR, O., LEIDNER, J. L., SONG, D. y SCHILDER, F. Interacting with financial data using natural language. En *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, páginas 1121–1124. 2016.
- PONNAMPERUMA, K., SIDDHARTHAN, A., ZENG, C., MELLISH, C. y VAN DER WAL, R. Tag2blog: Narrative generation from satellite tag data. En *Proceedings of the 51st annual meeting of the association for computational linguistics: System demonstrations*, páginas 169–174. 2013.
- PORTEL, F., REITER, E., GATT, A., HUNTER, J., SRIPADA, S., FREER, Y. y SYKES, C. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, vol. 173(7-8), páginas 789–816, 2009.
- POSTEL, J. B. Simple mail transfer protocol. Informe Técnico RFC 821, Information Sciences Institute, University of Southern California, 1982.
- RADICATI, S. y LEVENSTEIN, J. Email statistics report, 2015-2019. *The Radicati Group, INC., A Technology Market Research Firm, Palo Alto, CA, USA, Tech. Rep, February*, 2015.
- RADICATI, S. y LEVENSTEIN, J. Email statistics report, 2021-2025. *The Radicati Group, INC., A Technology Market Research Firm, Palo Alto, CA, USA, Tech. Rep, February*, 2021.
- RAMOS-SOTO, A., BUGARIN, A. J., BARRO, S. y TABOADA, J. Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data. *IEEE Transactions on Fuzzy Systems*, vol. 23(1), páginas 44–57, 2014.
- REITER, E. y DALE, R. *Building Natural Language Generation Systems*. Cambridge University Press, 2000. ISBN 9780521620369.
- REITER, E., MELLISH, C. y LEVINE, J. Automatic generation of technical documentation. *Applied Artificial Intelligence an International Journal*, vol. 9(3), páginas 259–287, 1995.
- REITER, E., ROBERTSON, R. y OSMAN, L. Types of knowledge required to personalise smoking cessation letters. En *Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making*, páginas 389–399. Springer, 1999.
- REITER, E., SRIPADA, S., HUNTER, J., YU, J. y DAVY, I. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, vol. 167(1-2), páginas 137–169, 2005.
- RESNICK, P. Internet message format. Informe Técnico RFC 5322, Qualcomm Incorporated, 2008.
- REYNOLDS, J. K. Post office protocol. Informe Técnico RFC 918, Information Sciences Institute, 1984.
- SEGAL, E. Survey finds email fatigue could lead 38 % of workers to quit their jobs. *Forbes*, 2021.

- SIDDHARTHAN, A. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, vol. 165(2), páginas 259–298, 2014.
- SIDDHARTHAN, A., GREEN, M. J., VAN DEEMTER, K., MELLISH, C. S. y VAN DER WAL, R. Blogging birds: Generating narratives about reintroduced species to promote public engagement. En *Proceedings of the 7th International Natural Language Generation Conference (INLG 2012)*. ACL Anthology, 2012.
- DEL SOCORRO BERNARDOS, M. ¿qué es la generación de lenguaje natural? una visión general sobre el proceso de generación. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, vol. 11(34), páginas 105–128, 2007.
- STOCK, O., ZANCANARO, M., BUSETTA, P., CALLAWAY, C., KRÜGER, A., KRUPPA, M., KUFLIK, T., NOT, E. y ROCCHI, C. Adaptive, intelligent presentation of information for the museum visitor in peach. *User Modeling and User-Adapted Interaction*, vol. 17(3), páginas 257–304, 2007.
- TANG, J., YANG, Y., CARTON, S., ZHANG, M. y MEI, Q. Context-aware natural language generation with recurrent neural networks. *arXiv preprint arXiv:1611.09900*, 2016.
- THEUNE, M., KLABBERS, E., DE PIJPER, J.-R., KRAHMER, E. y ODIJK, J. From data to speech: a general approach. *Natural Language Engineering*, vol. 7(1), páginas 47–86, 2001.
- THOMASON, J., VENUGOPALAN, S., GUADARRAMA, S., SAENKO, K. y MOONEY, R. Integrating language and vision to generate natural language descriptions of videos in the wild. En *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, páginas 1218–1227. 2014.
- TROOST, R., DORNER, S. y MOORE, K. Communicating presentation information in internet messages: The content-disposition header field. Informe Técnico RFC 2183, Internet Engineering Task Force (IETF), 1997.
- TURNER, R., SRIPADA, S., REITER, E. y DAVY, I. P. Selecting the content of textual descriptions of geographically located events in spatio-temporal weather data. En *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, páginas 75–88. Springer, 2007.
- VASWANI, A., SHAZER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł. y POLOSUKHIN, I. Attention is all you need. En *Advances in neural information processing systems*, páginas 5998–6008. 2017.
- VICENTE, M., BARROS, C., PEREGRINO, F. S., AGULLÓ, F. y LLORET, E. La generación de lenguaje natural: análisis del estado actual. *Computación y Sistemas*, vol. 19(4), páginas 721–756, 2015.
- VAN DER WAL, R., SHARMA, N., MELLISH, C., ROBINSON, A. y SIDDHARTHAN, A. The role of automated feedback in training and retaining biological recorders for citizen science. *Conservation Biology*, vol. 30(3), páginas 550–561, 2016.
- WANNER, L., BOSCH, H., BOUAYAD-AGHA, N., CASAMAYOR, G., ERTL, T., HILBRING, D., JOHANSSON, L., KARATZAS, K., KARPPINEN, A., KOMPATSIARIS, I. ET AL. Getting the environmental information across: from the web to the user. *Expert Systems*, vol. 32(3), páginas 405–432, 2015.

YU, J., REITER, E., HUNTER, J. y MELLISH, C. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, vol. 13(1), páginas 25–49, 2007.

Apéndice A

Título del Apéndice A

Contenido del apéndice

Apéndice **B**

Título del Apéndice B

Este texto se puede encontrar en el fichero Cascaras/fin.tex. Si deseas eliminarlo, basta con comentar la línea correspondiente al final del fichero TFMTeXiS.tex.

*-¿Qué te parece desto, Sancho? – Dijo Don Quijote –
Bien podrán los encantadores quitarme la ventura,
pero el esfuerzo y el ánimo, será imposible.*

*Segunda parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes*

*-Buena está – dijo Sancho –; fírmela vuestra merced.
–No es menester firmarla – dijo Don Quijote–,
sino solamente poner mi rúbrica.*

*Primera parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes*

