eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Grado en Ingeniería Informática

Computación

Trabajo de Fin de Grado

Adquisición de conocimiento léxico a partir de diccionarios

Autora

Leire Varela Aranguren

2021

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Grado en Ingeniería Informática

Computación

Trabajo de Fin de Grado

Adquisición de conocimiento léxico a partir de diccionarios

Autora

Leire Varela Aranguren

Director

German Rigau

Resumen

Este documento describe el Trabajo de Fin de Grado (TFG) de la estudiante de Ingeniería Informática Leire Varela. En él se trata el área del **Procesamiento del Lenguaje Natural** (PLN), más conocido como *Natural Language Processing* o NLP. El Procesamiento del Lenguaje Natural es una rama de la inteligencia artificial que se ocupa de la interacción entre los ordenadores y los seres humanos usando el lenguaje natural. La mayoría de las técnicas de PLN se basan en el aprendizaje automático para ser capaz de comprender y dar sentido al significado de los diferentes idiomas que existen.

En este TFG aplicaremos el PLN a través de la adquisición de conocimiento léxico a partir del diccionario en línea de Oxford. Explicaremos cómo y para qué hemos obtenido el diccionario completo con sus definiciones, dominios y ejemplos, y aplicaremos y evaluaremos un clasificador de dominios con modelos de lenguaje pre-entrenados.

El proyecto ha sido dirigido por German Rigau.

Índice general

Índice de figuras

Índice de tablas

1. CAPÍTULO

Introducción

1.1. Contexto

Este proyecto se enmarca dentro del área del Procesamiento del Lenguaje Natural. En concreto, en la adquisición de conocimiento léxico que contienen los diccionarios de uso común. Los diccionarios son una fuente muy rica de conocimiento léxico. En particular, trabajaremos con el diccionario Oxford[1], ya que destaca por la abundancia de frases de ejemplo que contiene. Tener a nuestra disposición varias frases de ejemplo en las que se pueden ver los diferentes usos de una palabra puede ser de gran ayuda en el área de la desambiguación lingüística, así como a la hora de generar diccionarios en otros idiomas haciendo uso de la traducción automática avanzada, o incluso para crear las definiciones de una palabra. Este diccionario también contiene dominios asignados a algunas definiciones. Un **dominio** adjudica un ámbito determinado a un significado.

El **lenguaje natural** es la forma más común y versátil que tiene la humanidad de transmitir información. Los humanos usamos el lenguaje, nuestro medio natural de comunicación, para codificar, almacenar, transmitir, compartir y manipular información. De hecho, la mayor parte de la información digital disponible es información no estructurada en forma de documentos (escritos o hablados) en múltiples lenguas, representando un desafío para toda organización pública o privada que quiera aprovechar su propia información. Hay muchos sistemas informáticos que procesan únicamente datos estructurados, por ejemplo, bases de datos con millones de registros y datos numéricos. Sin embargo, no

[1] https://www.lexico.com/

es trivial procesar la información digital no estructurada (texto y voz), ya que está sujeta a múltiples interpretaciones (ambigüedad), falta de conocimiento sobre el contexto y el mundo, y su complejidad intrínseca. [SEPLN, 2020]

El **Procesamiento del Lenguaje Natural**, abreviado PLN (*Natural Language Processing* o NLP en inglés) es un campo de las ciencias de la computación, de la inteligencia artificial y de la lingüística que estudia las interacciones entre las computadoras y el lenguaje humano. En particular, estudia cómo programar las computadoras para procesar y analizar largas cantidades de datos del lenguaje natural. Su objetivo es entender los contenidos de documentos, incluidos los matices contextuales que se pueden encontrar. [Wikipedia, 2021b]

Una forma de aplicar esta tecnología es extrayendo y almacenando conocimiento léxico de páginas web mediante la técnica de ***web scraping***. Con este método se pueden recolectar datos desde la web ya sea de forma manual o automática. En concreto, la palabra *scraping* nos da una intuición de lo que implica esta técnica, ya que traducida al castellano significa raspar, reunir o arañar.

En lingüística, la **semántica** es el estudio del significado de las palabras, construcciones, expresiones y significado de relaciones con las conexiones entre los usos de las palabras. Concretamente, el **conocimiento léxico** estudia el significado de una o un grupo de palabras (oraciones o unidades más grandes) [Castillo and Rigau, 2013].

En el ámbito de la lingüística computacional, la **desambiguación** del significado de las palabras es un problema abierto de PLN, que incluye el proceso de identificar con qué sentido se usa una palabra en los términos de una oración, o cuándo la palabra en cuestión tiene polisemia, es decir, una pluralidad de significados.

La **traducción automática** o *machine translation* en inglés es un área de la lingüística computacional que investiga el uso de software para traducir texto o habla de un lenguaje natural a otro. En su vertiente más básica, simplemente sustituye las palabras de un idioma por las de otro, pero es obvio que este procedimiento rara vez da lugar a una traducción buena, pues el léxico cambia en las distintas lenguas. [Wikipedia, 2021c]

1.2. Motivación

Los corpus de los diccionarios suelen estar disponibles en línea en formato electrónico. Sin embargo, a menudo les faltan oraciones de ejemplo. Por lo que hemos visto, el diccio-

nario en línea de Oxford es el único que contiene una cantidad abundante de oraciones de ejemplo, y esta ha sido la principal razón por la que nos ha parecido interesante trabajar con este diccionario en concreto.

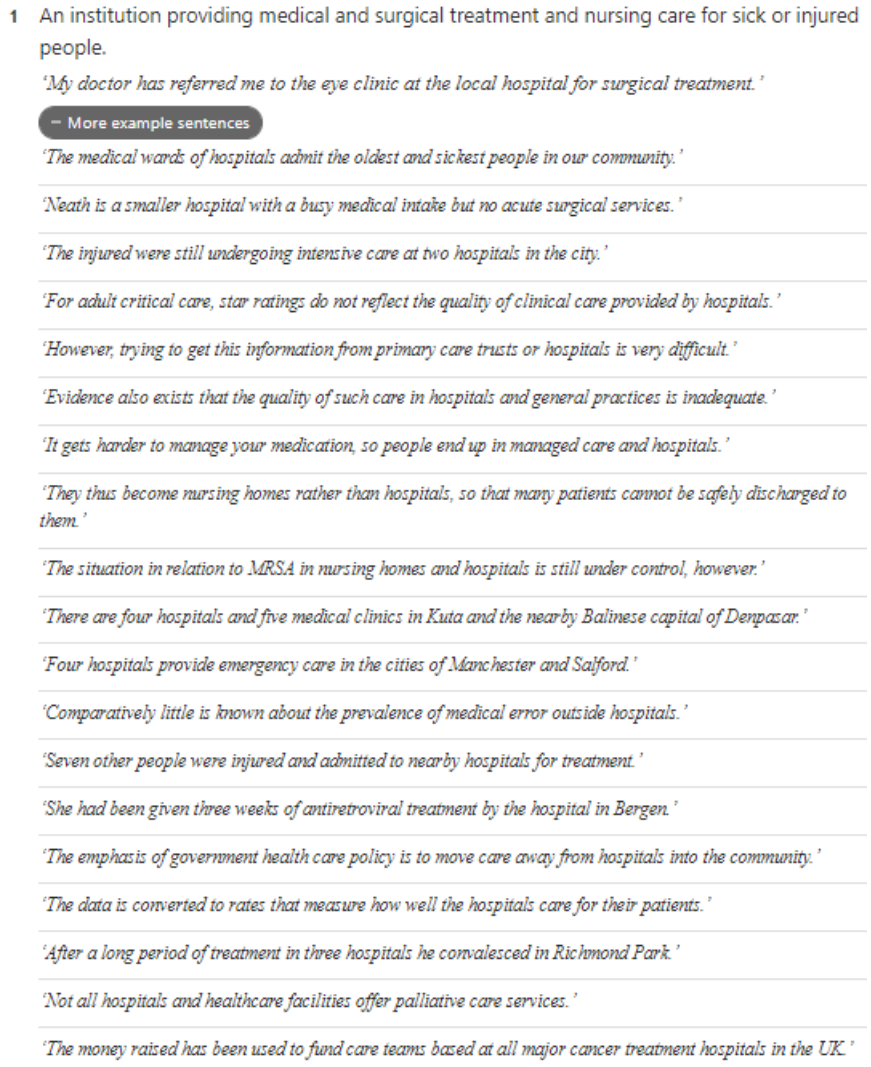


Figura 1.1: Definición y ejemplos de la entrada *hospital*

Además, contiene una alta variedad de dominios y muchas de sus definiciones están ligadas a alguno de ellos.

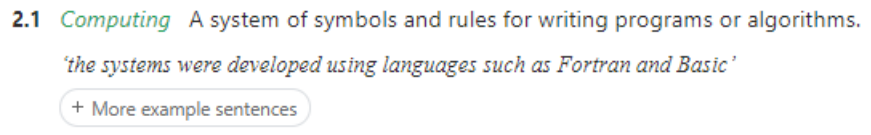


Figura 1.2: Subdefinición y dominio (en verde) de la palabra *language*

Por lo tanto, este proyecto se centrará en la adquisición de conocimiento léxico del diccionario de Oxford con el fin de hacer una aportación en el área del Procesamiento del Lenguaje Natural.

1.3. Objetivos

El objetivo principal de este proyecto es la adquisición de conocimiento léxico a partir del diccionario en línea de Oxford mediante la técnica de *web scraping*. Para ello, hemos usado y mejorado el código del proyecto xSense [Chang et al., 2018] que se encarga de adquirir las definiciones y ejemplos de este mismo diccionario.

Una vez obtenidas todas las definiciones, ejemplos, sinónimos y demás atributos de todas las palabras del diccionario, nuestro siguiente objetivo es agrupar los dominios que algunas definiciones tienen asignados. Estos **dominios** indican la pertenencia a una categoría determinada de una palabra. Tendremos que ver cuántos hay en el diccionario y cuáles son los más frecuentes.

Después probaremos el clasificador de dominios zero-shot con modelos de lenguaje preentrenados Ask2Transformers [Sainz and Rigau, 2020] desarrollado por Oscar Sainz y German Rigau que anota automáticamente los datos textuales sin ninguna supervisión. Dado un conjunto particular de etiquetas, el sistema tiene que clasificar los datos sin ejemplos previos. Nosotros lo probaremos con definiciones de Oxford y nuestro conjunto de etiquetas serán sus dominios.

Por último, nuestra tarea final es realizar distintas pruebas con Ask2Transformers utilizando distintos datos y conjuntos de etiquetas y comparar los resultados con la clasificación de dominios del diccionario de Oxford. Nuestro objetivo será obtener los mejores resultados posibles.

1.4. Estructura del documento

El documento está dividido en seis capítulos de la siguiente manera:

- **Capítulo 1**. Breve **introducción** del TFG. Ofrece el contexto del proyecto, la motivación para realizarlo, los objetivos principales y la estructura de este documento.

- **Capítulo 2**. Contiene la **planificación** y **gestión** del proyecto. Incluye un diagrama de Gantt.

- **Capítulo 3**. Visión general del **estado del arte** en la adquisición de conocimiento léxico para el Procesamiento Natural del Lenguaje. Se habla sobre otros trabajos previos que también han trabajado con el diccionario Oxford y que han influido en este proyecto; y de algunos modelos pre-entrenados del lenguaje como Ask2Transformers, HuggingFace y BERT.

- **Capítulo 4**. **Implementación** y **resultados** de obtener el diccionario en línea de Oxford completo con todos sus ejemplos y demás atributos. Está dividido en cuatro secciones principales. La primera y la última están dedicadas a una introducción (4.1) y las conclusiones finales (4.4). La segunda describe la implementación de dos programas realizados: obtención de los listados de palabras (4.2.1), y obtención del diccionario completo (4.2.2). La tercera sección muestra los resultados de estos dos programas (4.3.1 y 4.3.2).

- **Capítulo 5**. **Desarrollo** y **resultados** de obtener los dominios de Oxford y aplicar y evaluar el clasificador de dominios Ask2Transformers. Está dividido en cuatro secciones principales. Al igual que en el capítulo 4, la primera y última sección están dedicadas a la introducción (5.1) y a las conclusiones (5.4). La segunda habla sobre los dominios de Oxford y cómo los hemos obtenido y ordenado (5.2.1), y sobre el funcionamiento del programa Ask2Transformers y cómo lo hemos utilizado nosotros (5.2.2). La tercera sección muestra los resultados de cuatro pruebas realizadas en Ask2Transformers: dominios en mayúsculas y en minúsculas (5.3.1), 100 definiciones y 90 dominios (5.3.2), agrupación de dominios (5.3.3), dominios más frecuentes (5.3.4) y etiquetado de ejemplos (5.3.5).

- **Capítulo 6**. Principales **conclusiones** del proyecto y **futuras tareas** que pueden realizarse.

2. CAPÍTULO

Planificación y gestión del proyecto

2.1. Planificación del proyecto

En esta sección explicaremos brevemente cómo ha sido la planificación del proyecto. La planificación constituye un elemento fundamental, debido a que a través de ella se concretan los criterios y objetivos fundamentales del trabajo y se define la metodología que servirá de base para su realización. Dicha metodología debe incluir además una planificación temporal del desarrollo de cada fase del trabajo, de forma que se garantice la compatibilidad entre las distintas fases y que se asegure su terminación en el plazo establecido.

El trabajo se ha planificado en tres fases principales que se describen a continuación.

2.1.1. Introducción

Esta fase supone la primera toma de contacto con el Trabajo de Fin de Grado. Engloba el conocer el contexto y los objetivos del TFG, buscar una motivación, buscar y entender trabajos previos realizados por otros autores con el objetivo de tener una visión general inicial del proyecto; y el conocimiento de los modelos pre-entrenados del lenguaje para entender en qué consiste el mismo. Todo esto está descrito en los capítulos 1 y 3 de este documento.

2.1.2. Implementación y obtención del diccionario

En la fase de implementación y obtención del diccionario se incluye el código generado para la obtención de los listados de palabras y del contenido del diccionario completo. Contiene igualmente el análisis de los resultados obtenidos tanto de los listados de palabras como del diccionario completo y la elaboración de conclusiones.

2.1.3. Clasificación de dominios sin entrenamiento

En la tercera fase, se explica el desarrollo realizado para saber cuántos dominios contiene el diccionario y cuáles son los más frecuentes. A continuación se realizan distintas pruebas con el programa Ask2Transformers y se evalúan los resultados de este último junto con la clasificación creada por el diccionario en línea. Para poder llevar a cabo esta ejecución, se genera un código que se encargue de hacer las comparaciones y decir cuál es el porcentaje de acierto y cuál el de fallo.

2.2. Gestión del proyecto

Para la óptima gestión del proyecto, se ha realizado un seguimiento y control del mismo periódicamente. Se ha redactado esta memoria que describe todo lo que engloba el proyecto, se ha diseñado un póster que presenta de manera esquemática el TFG, y por último se ha hecho una presentación en formato de diapositivas para presentar el trabajo ante un tribunal.

En este apartado describiremos estos elementos necesarios para la gestión junto con un diagrama de Gantt al final.

2.2.1. Seguimiento y control

El seguimiento y control del proyecto se ha realizado mediante reuniones periódicas como a través del contacto mediante el correo electrónico. El objetivo principal de las reuniones mantenidas entre el director y la autora del trabajo ha consistido en la realización de un seguimiento del desarrollo del trabajo a lo largo de sus diferentes fases, de forma que se garantice la adecuación entre los objetivos inicialmente previstos y los contenidos que

se iban incorporando al documento. Las reuniones mantenidas han sido tanto de carácter presencial como telemático.

2.2.2. Memoria

A través de la Memoria se exponen los criterios y objetivos que han servido de base para el desarrollo del Trabajo Fin de Grado, se describen los contenidos de los distintos capítulos y se exponen los resultados y las conclusiones alcanzadas.

2.2.3. Póster

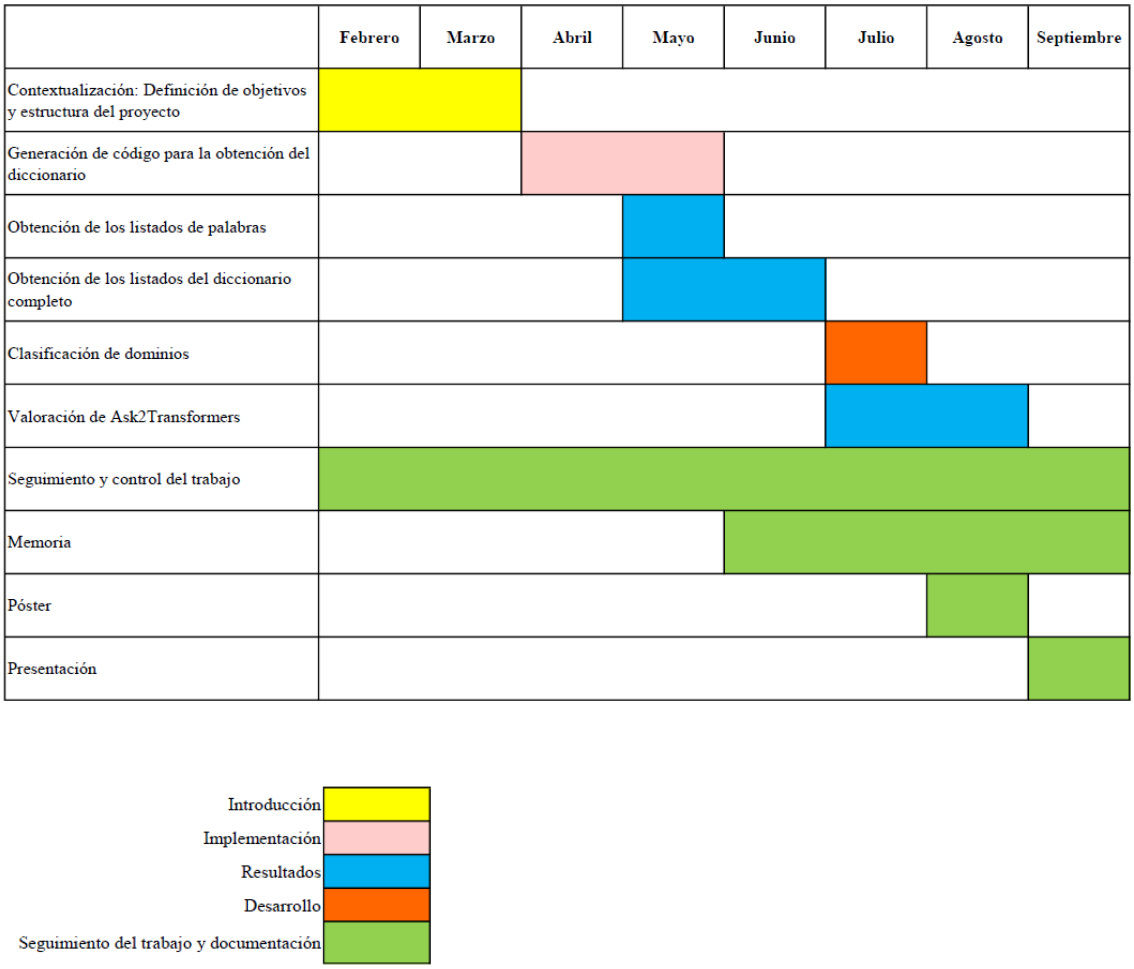
El Poster constituye un documento que pretende sintetizar, de forma gráfica y en una sola imagen, el contenido del trabajo realizado.

2.2.4. Presentación

La presentación del trabajo elaborado ante el tribunal se realiza en forma de diapositivas que pretenden condensar los principales contenidos del trabajo realizado. En su contenido se incluye documentación tanto gráfica como escrita con el objetivo de que permita una explicación más ilustrativa del trabajo desarrollado. Constituye un documento fundamental para conseguir transmitir de forma clara y concisa el objeto del trabajo, su contenido y alcance.

2.2.5. Diagrama de Gantt

A continuación se adjunta un Diagrama de Gantt en el que se resume el alcance de las distintas fases en que se ha estructurado la planificación del proyecto.


Febrero
Marzo
Abril
Mayo
Junio
Julio
Agosto
Septiembre
Contextualización: Definición de objetivos y estructura del proyecto
Generación de código para la obtención del diccionario
Obtención de los listados de palabras
Obtención de los listados del diccionario completo
Clasificación de dominios
Valoración de Ask2Transformers
Seguimiento y control del trabajo
Memoria
Póster
Presentación
Introducción
Implementación
Resultados
Desarrollo
Seguimiento del trabajo y documentación

3. CAPÍTULO

Estado del arte

3.1. Adquisición de conocimiento léxico usando diccionarios

La adquisición automática del conocimiento a partir de las fuentes disponibles, en su mayor parte corpus textuales, es una de las tareas más difíciles en PLN ya que requiere de alguna capacidad para entender el texto. Hace años fue un tema popular el diccionario legible por máquina o ***Machine-readable dictionary*** (MRD). Este diccionario se caracteriza por estar almacenado como datos de máquina (computadora) en lugar de impreso en papel. Al estar en forma electrónica, se puede cargar en una base de datos y consultar a través de una aplicación software [Wikipedia contributors, 2021b].

Más tarde con la aparición de **WordNet**[1] empezaron nuevos propósitos en sistemas de información que incluían desambiguación del significado de palabras, recuperación de información, clasificación automática de texto, resumen automático de texto, traducción automática e incluso generación de crucigramas [Wikipedia contributors, 2021c]. WordNet es una gran base de datos léxica del inglés. Todos los sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos cognitivos llamados *synsets*, cada uno expresando un concepto distinto. Se asemeja superficialmente a un tesauro, ya que agrupa las palabras en función de sus significados. Su objetivo era desarrollar un sistema que fuera consistente con el conocimiento adquirido a través de los años sobre cómo los humanos procesan el lenguaje. Por ejemplo, los individuos pueden verificar rápidamente que los canarios pueden volar porque un canario es un ave, pero cuesta más trabajo identificar

[1] http://wordnetweb.princeton.edu/perl/webwn

que un canario tiene piel. Esto sugiere que nosotros también almacenamos información semántica en una forma muy parecida a como lo hace WordNet, porque solo retenemos la información más específica que necesitamos para diferenciar un concepto en particular de otros conceptos similares [A. and R, 1972].

Antes de existir WordNet, el diccionario más usado era el ***Longman Dictionary of Contemporary English*** [2] (LDOCE). Este diccionario de aprendizaje avanzado proporciona definiciones mediante el uso de un vocabulario restringido, ayudando a los hablantes no nativos del inglés a comprender los significados fácilmente. Hoy en día está disponible tanto en línea como en papel [Wikipedia contributors, 2021a].

3.2. Trabajos previos

Uno de los objetivos principales de este proyecto es la adquisición de conocimiento léxico de un diccionario en línea. El proyecto **xSense** [Chang et al., 2018] obtiene las definiciones y ejemplos del diccionario Oxford mediante la técnica de *crawling*. Su objetivo es resolver tres principales problemas que encuentran los humanos a la hora de interpretar los *word embeddings*. ***Word embedding*** es el nombre de un conjunto de lenguajes de modelado y técnicas de aprendizaje en procesamiento del lenguaje natural en dónde las palabras o frases del lenguaje natural son representadas como vectores de números reales [Wikipedia, 2020]. A continuación mostramos los tres principales problemas que encuentran los humanos a la hora de interpretarlos:

1. **Polisemia**: Los *word embeddings* mezclan diferentes significados en un solo vector, que también se conoce como polisemia.

2. **Entender las dimensiones**: Los valores más altos y más bajos en las dimensiones de un vector de *embeddings* son difíciles de interpretar y analizar para un humano.

3. **Análisis semántico**: Solo podemos comprobar indirectamente los vecinos más cercanos para inspeccionar el significado semántico de un *word embedding*.

Este proyecto está interesado concretamente en el diccionario en línea de Oxford debido a que es el único que contiene una amplia variedad de frases de ejemplo por definición. [3]

[2] https://www.ldoceonline.com/
[3] https://github.com/MiuLab/xSense

A su vez, el proyecto xSense se basa en otro anterior [Gadetsky and Vetrov, 2018]. Este proyecto hace lo mismo, sin embargo, su conjunto de datos no contiene información completa disponible en línea, lo que dificulta su uso para diversas tareas.

3.3. Modelos pre-entrenados del lenguaje

Un modelo previamente entrenado es un modelo creado y entrenado por otra persona para resolver un problema. En la práctica, casi siempre esa persona es un gigante tecnológico o un grupo de investigadores estrella. Por lo general, eligen un conjunto de datos muy grande como su conjunto de datos base, como ImageNet[4] o Wikipedia Corpus[5]. Luego, crean una gran red neuronal para resolver un problema particular. Por supuesto, este modelo previamente entrenado debe hacerse público para que podamos tomarlo y reutilizarlo.

3.3.1. Ask2Transformers

El clasificador de dominios Zero-Shot con modelos de lenguaje pre-entrenados llamado **Ask2Transformers** [Sainz and Rigau, 2020] y desarrollado por Oscar Sainz y German Rigau es un sistema que explota diferentes modelos de lenguaje previamente entrenados para asignar etiquetas de dominio a los *synsets* de WordNet sin ningún tipo de supervisión. Además, el sistema no está restringido a utilizar un conjunto particular de etiquetas de dominio. Explotamos el conocimiento codificado dentro de diferentes modelos de lenguaje pre-entrenados listos para usar y formulaciones de tareas para inferir la etiqueta de dominio de una definición particular de WordNet. El sistema *zero-shot* propuesto logra un nuevo estado-del-arte en el conjunto de datos en inglés utilizado en la evaluación.[6]

3.3.2. HuggingFace

HuggingFace es una plataforma de alojamiento donde los usuarios pueden crear, comparar, compartir, versionar e implementar repositorios que pueden incluir modelos, conjuntos de datos y aplicaciones de aprendizaje automático. Proporciona distintos modelos de traducción a múltiples idiomas usando modelos llamados *transformers* que se utilizan para resolver todo tipo de tareas de PLN, como por ejemplo:

[4] https://www.image-net.org/
[5] https://es.wikipedia.org/wiki/Wikipedia:Portada
[6] https://github.com/osainz59/Ask2Transformers

- **Clasificar oraciones enteras:** Obtener el sentimiento de una revisión, detectar si un correo electrónico es *spam*, determinar si una oración es gramaticalmente correcta o si dos oraciones están lógicamente relacionadas o no.
- **Clasificar cada palabra en una oración:** Identificar los componentes gramaticales de una oración (sustantivo, verbo, adjetivo) o las entidades nombradas (persona, ubicación, organización).
- **Generación de contenido de texto:** Completar un mensaje con texto generado automáticamente, rellenar los espacios en blanco de un texto con palabras enmascaradas.
- **Extracción de una respuesta de un texto:** Dada una pregunta y un contexto, extraer la respuesta a la pregunta en función de la información proporcionada en el contexto.
- **Generar una nueva oración a partir de un texto de entrada:** Traducir un texto a otro idioma, resumir un texto.

Una tarea más desafiante es cuando necesitamos clasificar textos que no han sido etiquetados. Este es un escenario común en proyectos del mundo real porque anotar texto suele llevar mucho tiempo y requiere experiencia en el dominio. Para este caso de uso, el *pipeline* de clasificación ***zero-shot-classification*** es muy potente: le permite especificar qué etiquetas usar para la clasificación, por lo que no tiene que depender de las etiquetas del modelo entrenado previamente. Ya ha visto cómo el modelo puede clasificar una oración como positiva o negativa usando esas dos etiquetas, pero también puede clasificar el texto usando cualquier otro conjunto de etiquetas que desee.

El objeto más básico de la biblioteca *Transformers* es el ***pipeline***. Conecta un modelo con sus pasos necesarios de preprocesamiento y postprocesamiento, lo que nos permite introducir directamente cualquier texto y obtener una respuesta inteligible. [7]

3.3.3. BERT

BERT (*Bidirectional Encoder Representations from Transformers*) o Representación de Codificador Bidireccional de Transformadores es una técnica basada en redes neuronales para el pre-entrenamiento del procesamiento del lenguaje natural (PLN) desarrollada por

[7] https://huggingface.co/course/chapter1/3?fw=pt

Google.[Wikipedia, 2021a] Un grupo de investigadores que trabajan en el lenguaje de la Inteligencia Artificial de Google publicó BERT recientemente y está causando un gran revuelo por sus resultados increíblemente precisos en varias tareas de programación en lenguaje natural como MNLI (inferencia en lenguaje natural), Escuadrón V1.1 (respuesta a preguntas), y varias otras.

La implementación del entrenamiento del *Transformer* es una razón significativa por la que la comunidad de aprendizaje de máquinas considera BERT una innovación técnica esencial. El modelo de lenguaje de BERT promete llevar el aprendizaje automático a nuevas alturas. Es opuesto a los esfuerzos anteriores que se centraban en secuencias de texto que comenzaban con un entrenamiento de derecha a izquierda o de izquierda a derecha.

Los resultados indican que los modelos de lenguaje entrenados bidireccionalmente tienen una profunda comprensión del flujo y el contexto del lenguaje en comparación con los modelos de lenguaje basados en una sola dirección.

4. CAPÍTULO

Obtención el diccionario

4.1. Introducción

En este capítulo, explicaremos cómo hemos obtenido y descargado en unos ficheros de texto todas las palabras con sus definiciones, frases de ejemplo, dominios y demás contenidos del diccionario en línea de Oxford.[1]

Para entender bien el proceso, primero hace falta saber en qué consiste la técnica de ***web scraping***. Esta técnica sirve para obtener información de sitios web mediante programas de software.

Este proyecto ha reusado y mejorado el código del proyecto **xSense** [Chang et al., 2018]. Entre otras tareas, dicho proyecto se encarga de obtener las definiciones y los ejemplos de la palabra que el usuario le pase como parámetro. Su código está abierto al público en la plataforma de Github.[2]

Todo el código ha sido desarrollado en **Python** haciendo uso de la herramienta **Google Colaboratory**.[3] El código está subido en la plataforma de **GitHub**[4] y tanto el código como los resultados obtenidos han sido guardados en **Google Drive**.[5]

[1] https://www.lexico.com/
[2] https://github.com/MiuLab/xSense
[3] https://colab.research.google.com/
[4] https://github.com/leirevaran/lexico
[5] https://drive.google.com/drive/folders/1T-dBDwEtl_IVs0OytPycFedGj07YahIF?usp=sharing

4.2. Implementación

Hemos implementado dos programas principales para poder obtener todo el diccionario. Antes de explicar el funcionamiento de ambos programas, cabe destacar que se ha hecho uso de la herramienta Beautiful Soup. **Beautiful Soup**[6] es una biblioteca que está diseñada para analizar documentos HTML y extraer su información. Por otro lado, para poder entender bien cómo se han implementado los programas, hay que explicar que hemos accedido a los ficheros fuente de los HTML de las diferentes páginas web con las que hemos trabajado usando el navegador Chrome haciendo click derecho en el ratón y seleccionando la opción *Inspeccionar* (o pulsando a la vez las teclas *Ctrl+Mayús+I)*. Esta opción abre una ventana en la parte derecha de la pantalla que contiene el documento fuente en HTML de la página web que estamos visitando. Dicha ventana nos muestra un documento que contiene todas las etiquetas y nombres de las clases que usaremos (y las que no). El acceso a esta información es pública.

A continuación, se explica el funcionamiento de ambos programas.

4.2.1. Obtención de los listados de palabras

El objetivo del primer programa es obtener e imprimir en un fichero de texto todas las palabras contenidas en el diccionario en línea de Oxford. La implementación es la siguiente:

- El programa accede a la URL `https://www.lexico.com/list/` añadiéndole al final la letra del abecedario que le interese al usuario, la cual pasará como parámetro. Esta dirección contiene principalmente el listado de palabras que comienzan por la letra que el usuario haya indicado. El programa lee esta página web en formato HTML gracias a la biblioteca Beautiful Soup. Acto seguido, elimina de la información obtenida aquello que no vamos a utilizar. Para hacer esto, hay que indicarle qué clase (o clases) del HTML no nos interesan.
- A continuación, el programa obtiene el número de páginas web que tiene la letra que estemos tratando. Este dato se encuentra en la paginación, al final del listado de palabras. El botón *Last* contiene el número de la última página, lo que sirve para saber a cuántas páginas hay que acceder para obtener el listado completo de palabras de una letra en concreto.

[6] `https://www.crummy.com/software/BeautifulSoup/`

Figura 4.1: Paginación de los listados de palabras

- El siguiente paso consiste en llamar a otra función tantas veces como número de páginas tenga la letra. A esta función se le pasan tres parámetros: la letra, el fichero donde se imprimirá el resultado, y el número de página a tratar.
- Esta segunda función accede a la URL https://www.lexico.com/list/ añadiéndole al final la letra, más '/', más el número de página. De nuevo, el programa lee esta página web en formato HTML gracias a la biblioteca Beautiful Soup y elimina de la información obtenida aquello que no vamos a utilizar.
- Por último, el programa accede a las entradas que contiene una página en concreto mediante las etiquetas del HTML, las enumera y las imprime. Cada línea del fichero en el que se imprime el resultado contiene la información de una palabra. Concretamente: la letra con la que empieza, el número de página, el número de la palabra y la palabra.

```
LETTER: p PAGE: 14 NUMBER: 229 WORD: party
```

Figura 4.2: Palabra *party* en el fichero resultante

Hay que remarcar que el número de página es orientativo, ya que el programa cuenta 300 palabras por página con la intención de que el fichero quede más ordenado. Sin embargo, esto no tiene por qué ir acorde al orden en el que están representadas las palabras en la página web, ya que ahí no siempre hay 300 palabras por página.

4.2.2. Obtención del diccionario completo

El segundo programa consiste en imprimir en un fichero de texto los dominios, las definiciones, los ejemplos y los sinónimos (entre otras cosas) de todas las palabras del diccionario en línea de Oxford. Para poder hacer uso de este programa, primero hay que obtener los ficheros devueltos en el programa explicado en el apartado anterior, es decir, los ficheros que contienen por cada letra todas las palabras del diccionario.

Este proceso lo hemos implementado de la siguiente manera:

- El usuario debe pasarle a la función dos parámetros: un fichero del que lee, y otro en el que escribe. El fichero del que lee es el que contiene el listado de palabras que comienzan por alguna letra del abecedario. Por cada línea: escribe la línea entera en el fichero de escritura, se queda con la palabra, la modifica para que no tenga tildes, espacios en blanco u otros caracteres especiales, y llama a otra función a la que le pasa como parámetros la palabra modificada y el fichero de escritura.
- Esta función accede a la URL https://en.oxforddictionaries.com/definition/ añadiéndole al final la palabra que ha recibido como parámetro (y es por eso que primero había que deshacerse de los espacios en blanco, tildes u otros caracteres especiales). Esta dirección contiene todas las definiciones y todo tipo de información de la palabra. Al igual que el programa mencionado anteriormente, lee esta página web en formato HTML haciendo uso de la biblioteca Beautiful Soup y elimina de la información obtenida aquello que no nos interesa.
- Haciendo uso de las etiquetas y clases del documento HTML de la página web, primero obtiene la información de la categoría morfosintáctica (si la palabra que estamos tratando es un sustantivo, un verbo, un adjetivo, etc.) y, en su caso, la palabra en plural, el infinitivo, etc.

```
WORD: do        POS: verb       (does,  doing,  did,  done)
```

Figura 4.3: Categoría morfosintáctica de la palabra *do* en el fichero resultante

- Después, obtiene la transitividad del verbo en caso de tenerla ([no object] / [with object]) e imprime una línea en el fichero que indica: la palabra que estamos tratando, el tipo de palabra que es, y, en caso de tenerla, la transitividad.

```
WORD: passage   POS: verb       [with object]
```

Figura 4.4: Transitividad de la palabra *passage* en el fichero resultante

- Lo siguiente que hace el programa es comprobar si la palabra tiene asignado algún dominio (Chemistry/Medicine/...). Normalmente, el dominio viene marcado en color verde. En caso afirmativo, imprime una línea en el fichero que indica: la palabra que estamos tratando, el tipo de palabra que es, y, en caso de tenerlo ya asignado, el dominio al que pertenece.

```
WORD: family    POS: noun       DOM: Biology
```

Figura 4.5: Dominio de la palabra *family* en el fichero resultante

- Luego, comprueba si la palabra tiene una definición o una referencia a otra palabra. Una definición es una frase que describe el significado de una palabra, mientras que la referencia es una palabra cuya definición debemos consultar si queremos saber la definición de la palabra que estamos tratando. En ambos casos, el programa imprime en el fichero de salida la palabra, el tipo de palabra que es y su definición (enumerada) o su referencia.

```
WORD: party     POS: noun       DEF: 1  A social gathering
```

Figura 4.6: Parte de la definición 1 de la palabra *party* en el fichero resultante

```
WORD: phallicism        POS: noun       REF: phallus
```

Figura 4.7: Referencia de la palabra *phallicism* en el fichero resultante

- Cada definición o cada referencia puede contener numerosos ejemplos. En caso de tenerlos, el programa los enumera y los imprime.

```
WORD: party     POS: noun       DEF: 1  EX: 1   'an engagement party'
WORD: party     POS: noun       DEF: 1  EX: 2   'Sometimes, caterers
WORD: party     POS: noun       DEF: 1  EX: 3   'The hotel staff enco
```

Figura 4.8: Parte de los ejemplos de la palabra *party* en el fichero resultante

- Al igual que con los ejemplos, cada definición puede contener una lista de sinónimos que se imprimirá siempre que exista.

```
WORD: family    POS: noun       DEF: 1  SYN: household, ménage
```

Figura 4.9: Sinónimos de la palabra *family* en el fichero resultante

- También puede haber subdefiniciones por cada definición. O sea, una o varias definiciones que derivan de una definición. Estas subdefiniciones se enumeran y se comprueba, del mismo modo que con las definiciones, si contienen ejemplos y/o sinónimos. Toda esta información también quedará plasmada en el fichero de salida.

```
WORD: family    POS: noun       DEF: 1  SUBDEF: 1       A group
```

Figura 4.10: Parte de una subdefinición de la palabra *family* en el fichero resultante

- Este proceso se repite por cada línea (o palabra) contenida en el fichero de entrada. Es importante saber que, entre palabra y palabra, el programa principal espera entre uno y dos segundos para que el diccionario en línea no reciba demasiadas llamadas de golpe y se bloquee el acceso.

En relación a este último punto, nos gustaría remarcar que una primera versión del programa no recibía ningún fichero de entrada con las palabras a procesar. En su lugar, para poder obtener las palabras del diccionario, llamaba al programa definido en el apartado 4.2.1. Esta opción hacía demasiadas llamadas seguidas a la página web oficial del diccionario: una para obtener la palabra, y otra para obtener la información de esa palabra. Es por eso que en la versión final primero obtenemos el léxico completo del diccionario y luego obtenemos la información de cada una de las palabras.

A continuación se muestra por partes un ejemplo con la palabra *party* de cómo queda reflejada en el fichero de salida la información del diccionario.

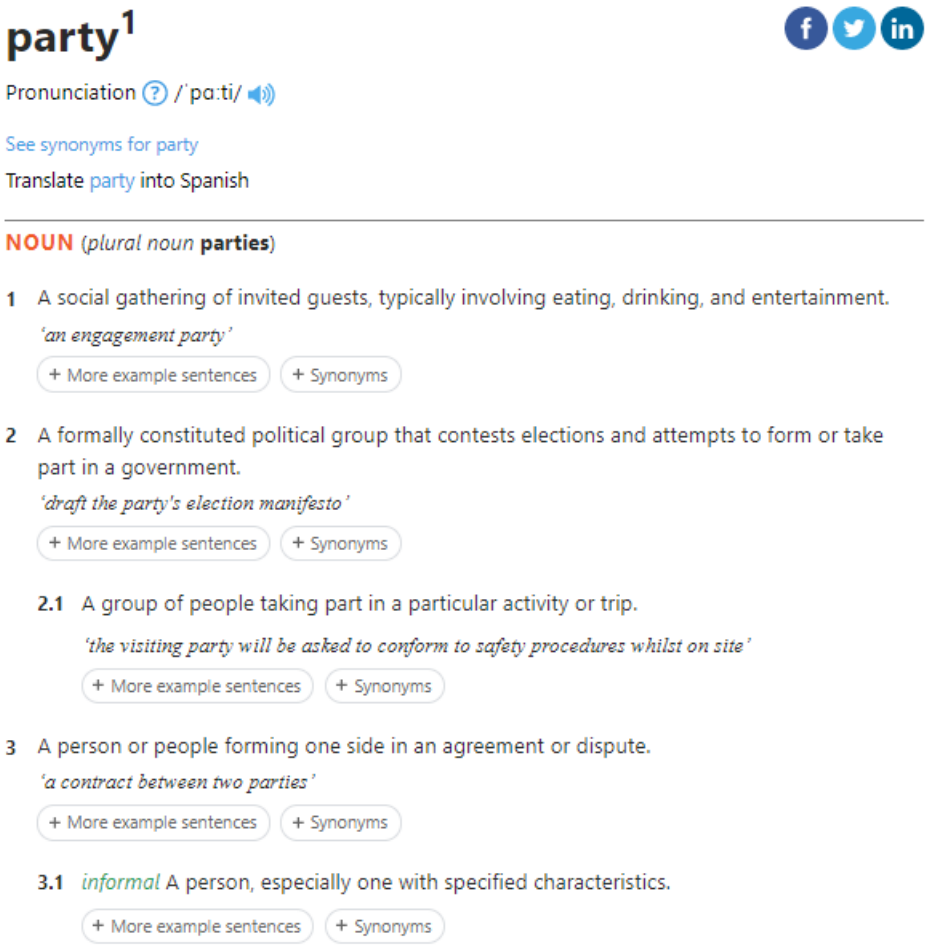


Figura 4.11: Definiciones de la palabra *party* cuando es un sustantivo

```
LETTER: p PAGE: 14 NUMBER: 229 WORD: party
WORD: party     POS: noun       (parties)
WORD: party     POS: noun       DEF: 1  A social gathering of invited guests, typically involving eating, drinking, and en
WORD: party     POS: noun       DEF: 1  EX: 1   'an engagement party'
WORD: party     POS: noun       DEF: 1  EX: 2   'Sometimes, caterers serving at parties and social gatherings order large
WORD: party     POS: noun       DEF: 1  EX: 3   'The hotel staff encourages the use of this area for social gatherings and
WORD: party     POS: noun       DEF: 1  EX: 4   'A buffet of finger foods is the perfect way to serve guests at an anniver
WORD: party     POS: noun       DEF: 1  EX: 5   'It was a common drink, brewed by 18th century farm owners at family parti
WORD: party     POS: noun       DEF: 1  EX: 6   'This was given out to guests at the party, but a few bottles were held ba
WORD: party     POS: noun       DEF: 1  EX: 7   'As the week turns to weekend, teenagers rush to the bottle shops to buy t
WORD: party     POS: noun       DEF: 1  EX: 8   'Day patients have been celebrating the festive season all week with speci
WORD: party     POS: noun       DEF: 1  EX: 9   'Avoid having many long holiday gatherings and parties with large numbers
WORD: party     POS: noun       DEF: 1  EX: 10  'She is looking up at the group and beaming at them, like someone at a dri
WORD: party     POS: noun       DEF: 1  EX: 11  'In the past, the youths usually ended the parade with a party, where they
WORD: party     POS: noun       DEF: 1  EX: 12  'Ana and I had discussed before the party what kind of drink we would be c
WORD: party     POS: noun       DEF: 1  EX: 13  'I cried for every birthday when no matter how many I invited to his party
WORD: party     POS: noun       DEF: 1  EX: 14  'To carry on with the theme of the party, let each guest make a list of se
WORD: party     POS: noun       DEF: 1  EX: 15  'A week ago on Saturday, my brother broke his toe while drunk at a party a
WORD: party     POS: noun       DEF: 1  EX: 16  'This is the first time I've done a summer holiday event, I usually do sch
WORD: party     POS: noun       DEF: 1  EX: 17  'Upstairs, the walls are decorated with photos of smiling people at partie
WORD: party     POS: noun       DEF: 1  EX: 18  'Other activities include a party to celebrate the club's first anniversar
WORD: party     POS: noun       DEF: 1  EX: 19  'It seemed that there was always something to do, be it orientation activi
WORD: party     POS: noun       DEF: 1  EX: 20  'Companies that are no longer in business spent millions on parties and pr
WORD: party     POS: noun       DEF: 1  EX: 21  'During the weeks preceding my graduation from high school several people
WORD: party     POS: noun       DEF: 1  SYN: social gathering, gathering, social occasion, social event, social function,

WORD: party     POS: noun       DEF: 2  A formally constituted political group that contests elections and attempts to for
WORD: party     POS: noun       DEF: 2  EX: 1   'draft the party's election manifesto'
WORD: party     POS: noun       DEF: 2  EX: 2   'Each ballot paper has a list of all registered political parties contesti
WORD: party     POS: noun       DEF: 2  EX: 3   'The Labour, Conservative and Liberal Democrat parties are contesting ever
WORD: party     POS: noun       DEF: 2  EX: 4   'It broke a 40-year monopoly of the two openly capitalist parties over wor
WORD: party     POS: noun       DEF: 2  EX: 5   'To win elections, politicians and parties wage costly campaigns.'
WORD: party     POS: noun       DEF: 2  EX: 6   'There was no attempt made by other parties to debate the issue.'
WORD: party     POS: noun       DEF: 2  EX: 7   'No wonder there is growing disillusionment with all mainstream parties an
WORD: party     POS: noun       DEF: 2  EX: 8   'At election times the party is dependent on resources and activists from
WORD: party     POS: noun       DEF: 2  EX: 9   'All constitutional parties opposed to the pact were unionist, and they ha
WORD: party     POS: noun       DEF: 2  EX: 10  'Both the ruling and opposition parties suspended all campaign activities
WORD: party     POS: noun       DEF: 2  EX: 11  'He remained respected in the party, in whose activities he took a close i
WORD: party     POS: noun       DEF: 2  EX: 12  'We should be able to build a broad movement which is not the product of a
WORD: party     POS: noun       DEF: 2  EX: 13  'The rally was organized by a newly powerful coalition of fundamentalist r
WORD: party     POS: noun       DEF: 2  EX: 14  'The prime minister could also seek smaller religious parties to bolster h
WORD: party     POS: noun       DEF: 2  EX: 15  'One is simply covering the events that happened, the campaign activities
WORD: party     POS: noun       DEF: 2  EX: 16  'In the following year, the ruling and opposition parties formed a coaliti
WORD: party     POS: noun       DEF: 2  EX: 17  'He promised to prepare the ground within his party, but his departure has
WORD: party     POS: noun       DEF: 2  EX: 18  'The new structure should operate under the jurisdiction of the Finance Mi
WORD: party     POS: noun       DEF: 2  EX: 19  'It is a party of working people against the Republican Party of corporati
WORD: party     POS: noun       DEF: 2  EX: 20  'If the ruling party doesn't perform well, the opposition can offer a viab
WORD: party     POS: noun       DEF: 2  EX: 21  'Across the entire party there is agreement - Labour has no chance of addi
WORD: party     POS: noun       DEF: 2  SYN: faction, political party, group, grouping, side, alliance, affiliation, assoc

WORD: party     POS: noun       DEF: 2  SUBDEF: 1       A group of people taking part in a particular activity or trip.
WORD: party     POS: noun       DEF: 2  SUBDEF: 1       EX: 1   'the visiting party will be asked to conform to safety pro
WORD: party     POS: noun       DEF: 2  SUBDEF: 1       EX: 2   'After an unsuccessful trip his hunting party bought him a
WORD: party     POS: noun       DEF: 2  SUBDEF: 1       EX: 3   'Private parties can book for trips along the coastline or
WORD: party     POS: noun       DEF: 2  SUBDEF: 1       EX: 4   'The most organised person in our party had brought a torc
WORD: party     POS: noun       DEF: 2  SUBDEF: 1       EX: 5   'There were 35 people on the tour and trouble flared when
WORD: party     POS: noun       DEF: 2  SUBDEF: 1       EX: 6   'Moving forward to the game's present day, you'll get to m

WORD: party     POS: noun       DEF: 3  A person or people forming one side in an agreement or dispute.
WORD: party     POS: noun       DEF: 3  EX: 1   'a contract between two parties'
WORD: party     POS: noun       DEF: 3  EX: 2   'In such cases, resort to binding adjudication will require the agreement
WORD: party     POS: noun       DEF: 3  EX: 3   'She accused both parties in the dispute of losing sight of the fact that
WORD: party     POS: noun       DEF: 3  EX: 4   'This will delay the much needed reforms as the various parties dispute th
WORD: party     POS: noun       DEF: 3  EX: 5   'There was an agreement between the parties under which the defendants wou
WORD: party     POS: noun       DEF: 3  EX: 6   'We consider a lease to be a private contractual agreement between two par
WORD: party     POS: noun       DEF: 3  EX: 7   'Conciliation officers will seek to resolve disputes by agreement between
WORD: party     POS: noun       DEF: 3  EX: 8   'I found the agreement eminently sensible, safeguarding the interests of p
WORD: party     POS: noun       DEF: 3  EX: 9   'It is very important to understand that the only settlement that will sur
WORD: party     POS: noun       DEF: 3  EX: 10  'But both parties are confident an agreement can be reached.'
WORD: party     POS: noun       DEF: 3  EX: 11  'The warring parties signed a ceasefire agreement on April 8 to would allo
WORD: party     POS: noun       DEF: 3  EX: 12  'Educational activities that benefit all parties are not impossible, but d
WORD: party     POS: noun       DEF: 3  EX: 13  'That is why the council is presently consulting with all interested parti
WORD: party     POS: noun       DEF: 3  EX: 14  'Attendance will be by invitation from the agency to organisations, intere
WORD: party     POS: noun       DEF: 3  EX: 15  'These were mutually exclusive areas of medical activity, as the parties a
WORD: party     POS: noun       DEF: 3  EX: 16  'Tensions rose when there was a perception among people that the two parti
WORD: party     POS: noun       DEF: 3  EX: 17  'So you know the phone lines between the two parties were burning up last
WORD: party     POS: noun       DEF: 3  EX: 18  'As you can see, there are no answers here, and the battle lines drawn by
WORD: party     POS: noun       DEF: 3  EX: 19  'All parties agree that the old legislation is not working and that someth
WORD: party     POS: noun       DEF: 3  EX: 20  'We would like to hear the views of parents and all other interested parti
WORD: party     POS: noun       DEF: 3  EX: 21  'The UN, the United States, Europe, and other interested parties urgently
WORD: party     POS: noun       DEF: 3  SYN: litigant, plaintiff, defendant

WORD: party     POS: noun       DEF: 3  SUBDEF: 1       A person, especially one with specified characteristics.
WORD: party     POS: noun       DEF: 3  SUBDEF: 1       EX: 1   'an old party has been coming in to clean'
WORD: party     POS: noun       DEF: 3  SUBDEF: 1       EX: 2   'The party on the line evidently had no idea what has happ
WORD: party     POS: noun       DEF: 3  SUBDEF: 1       EX: 3   'Seems it all began when an interested party dropped him a
WORD: party     POS: noun       DEF: 3  SUBDEF: 1       EX: 4   'A large proportion of money laundering activities involve
```

Figura 4.12: Parte del resultado de la palabra *party* cuando es un sustantivo

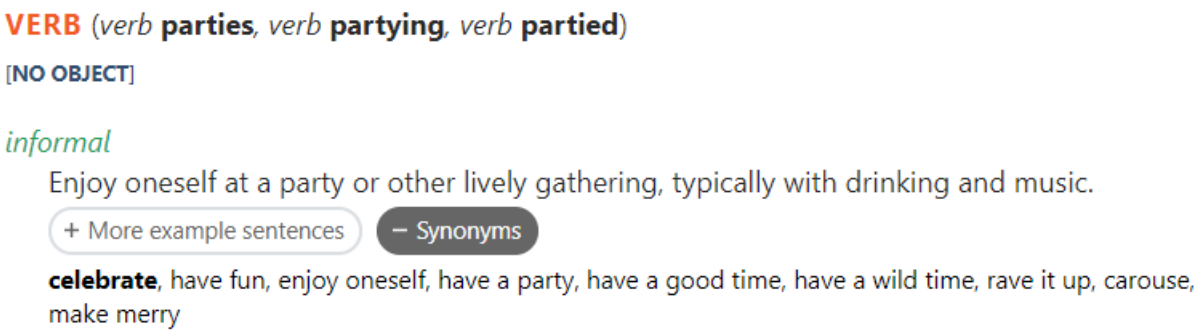


Figura 4.13: Definición de la palabra *party* cuando es un verbo

```
WORD: party     POS: verb       (parties,  partying,  partied) [no object]
WORD: party     POS: verb       DEF: 4  Enjoy oneself at a party or other l
WORD: party     POS: verb       DEF: 4  EX: 1   'put on your glad rags and
WORD: party     POS: verb       DEF: 4  EX: 2   'Maybe it's because we just
WORD: party     POS: verb       DEF: 4  EX: 3   'Everyone old and young brc
WORD: party     POS: verb       DEF: 4  EX: 4   'Some people just come for
WORD: party     POS: verb       DEF: 4  EX: 5   'A large number of family a
WORD: party     POS: verb       DEF: 4  EX: 6   'We partied into the night
WORD: party     POS: verb       DEF: 4  EX: 7   'Accordingly, she partied,
WORD: party     POS: verb       DEF: 4  EX: 8   'Oh, it's been a jolly time
WORD: party     POS: verb       DEF: 4  EX: 9   'For at least a small secti
WORD: party     POS: verb       DEF: 4  EX: 10  'She described her whole li
WORD: party     POS: verb       DEF: 4  EX: 11  'By the time college came a
WORD: party     POS: verb       DEF: 4  EX: 12  'I love partying in a safe
WORD: party     POS: verb       DEF: 4  EX: 13  'After cleaning himself up,
WORD: party     POS: verb       DEF: 4  EX: 14  'This is what it's like for
WORD: party     POS: verb       DEF: 4  EX: 15  'After dinner with multiple
WORD: party     POS: verb       DEF: 4  EX: 16  'That's not to say that I'm
WORD: party     POS: verb       DEF: 4  EX: 17  'I must be getting old, bec
WORD: party     POS: verb       DEF: 4  EX: 18  'Three years later he was p
WORD: party     POS: verb       DEF: 4  EX: 19  'The team had been partying
WORD: party     POS: verb       DEF: 4  EX: 20  'They are said to be workir
WORD: party     POS: verb       DEF: 4  EX: 21  'The real problem is that s
WORD: party     POS: verb       DEF: 4  SYN: celebrate, have fun, enjoy one
```

Figura 4.14: Parte del resultado de la palabra *party* cuando es un verbo

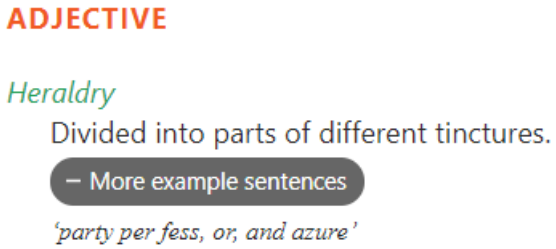


Figura 4.15: Definición de la palabra *party* cuando es un adjetivo

```
WORD: party     POS: adjective  DOM: Heraldry
WORD: party     POS: adjective  DEF: 5  Divided into parts of different tinctures.
WORD: party     POS: adjective  DEF: 5  EX: 1   'party per fess, or, and azure'
```

Figura 4.16: Resultado de la palabra *party* cuando es un adjetivo

4.3. Resultados

Tras ejecutar los programas implementados hemos ido obteniendo ficheros de texto con la información requerida. Tanto el programa que obtiene los listados de palabras como el que obtiene el diccionario completo, los hemos ejecutado (como mínimo) veintisiete veces: una por cada letra, de la *a* a la *z*, más los caracteres especiales, representados por el símbolo ♯ o el número *0*.

4.3.1. Obtención de los listados de palabras

Gracias a que el primer programa enumera las palabras que hay por cada letra, sabemos cuántas palabras contiene el diccionario en línea.

Letra	Cantidad palabras
0	567
a	13.538
b	11.397
c	14.485
d	8.315
e	6.663
f	8.678
g	8.094
h	8.462
i	6.731
j	1.949
k	2.453
l	7.549
m	19.607
n	8.229
o	8.920
p	31.543
q	1.696
r	16.468
s	23.868
t	10.698
u	5.947
v	2.606
w	6.777
x	143
y	906
z	734
Total	237.023

Tabla 4.1: Cantidad de palabras que contiene el diccionario Oxford

Como podemos ver en la tabla, la letra que más palabras contiene es la *p* con 31.543 palabras. En el otro extremo está la letra *x*, que con tan solo 143 palabras es la letra que menos palabras contiene.

4.3.2. Obtención del diccionario completo

Sacar toda la información de las 237.023 palabras del diccionario ha sido lo más costoso en cuanto a tiempo de ejecución. De media, en una hora de ejecución se imprimían cinco páginas con 300 palabras cada una, es decir, se imprimían alrededor de unas 1.500 pala-

bras por hora. Esto supone varios días de ejecución hasta poder imprimir la información de todas las palabras del diccionario. Además, se realizaron varias pruebas iniciales con distintas letras hasta que logramos imprimir casi todas las palabras. Decimos *casi todas* porque los ficheros obtenidos no contienen la información de absolutamente todas las palabras del diccionario, pero sí de casi todas. A veces, esto es debido a que la entrada contiene algún carácter especial, como es el caso de *æ*.
En esto influyó mucho el poner un tiempo de espera de entre uno y dos segundos entre palabra y palabra, ya que al principio no estaban estas líneas de código y, en cuanto estuvieron, el programa empezó a fallar menos en la impresión.

Si sumamos todas las definiciones (sin contar subdefiniciones) y todas las frases de ejemplo que contienen las palabras de todo el diccionario (incluidas las de las subdefiniciones), obtenemos los siguientes datos.

	Cantidad definiciones	Cantidad ejemplos
Total	314719	5961339
Media	1,33	25,15

Tabla 4.2: Suma y media de definiciones y ejemplos

Es decir, un total de 1,33 definiciones y 25,15 frases de ejemplo de media por cada palabra. Esto quiere decir que hay muchas palabras monosémicas (de un solo sentido).

Además, también hemos contado la cantidad de veces que una definición o una subdefinición contienen sinónimos: 315852 veces.

4.3.3. Comparación con otros proyectos

Con la información obtenida en las pruebas realizadas, podemos comparar los resultados con el proyecto **xSense** [Chang et al., 2018]. Este proyecto ya hacía una comparación de sus resultados con el trabajo de [Gadetsky and Vetrov, 2018].

La siguiente tabla, por lo tanto, muestra la comparación entre los dos proyectos anteriores ya mencionada, con la diferencia de que le hemos añadido una tercera columna al final con los datos obtenidos por nuestro proyecto.

Atributo	Otros	xSense	Nuestro
Cantidad palabras	36767	31798	237023
Cantidad ejemplos por definición de media	1	27	18,94

Tabla 4.3: Comparación de resultados con otros trabajos

La media de cantidad de frases de ejemplo por definición de nuestro proyecto se ha obtenido de los datos representados en la tabla 4.2. Se ha hecho una división entre la cantidad de ejemplos y la cantidad de definiciones que hay por palabra: $5961339 \div 314719 = 18,94$.

Como se puede observar, nosotros hemos obtenido una cantidad de palabras seis veces mayor que el proyecto de Gadetsky, Yakubovskiy y Vetrov, y siete veces mayor que el proyecto xSense. A su vez, nuestro proyecto obtiene más frases de ejemplo que el primer proyecto y menos que el segundo.

Pero, ¿por qué hay tanta diferencia entre los resultados obtenidos por nuestro proyecto y el del resto?

Pues bien, en cuanto a la cantidad de palabras, una razón puede ser que haya tres años de diferencia entre los proyectos anteriores y el nuestro. Esto quiere decir que quizá ahora el diccionario contenga más palabras que en 2018.
En cuanto a la media de cantidad de ejemplos por definición, es normal que haya diferencias porque este dato se ve directamente alterado según la cantidad de palabras y de definiciones que considere el proyecto. Aun así, puede ser que al primero no le interesaran tanto las frases de ejemplo y por eso sacara solo una por definición.

Otras diferencias entre el proyecto xSense y el nuestro son: 1) que ellos no obtienen las subdefiniciones (aunque sí sus ejemplos), 2) que nosotros obtenemos los sinónimos y ellos no, 3) nosotros obtenemos también los dominios, 4) algunas características de la categoría morfosintáctica en caso de tenerlas (como por ejemplo, el infinitivo) y 5) la transitividad. Además, toda la información la imprimimos de una forma clara y estructurada para que sea más fácil acceder a ella.

4.4. Conclusiones

Después de haber realizado el trabajo descrito en este capítulo, se pueden sacar las siguientes conclusiones:

1. Vista la cantidad de ejemplos que contiene el diccionario en línea de Oxford, podemos confirmar que este es un buen recurso para saber y entender mejor el significado de las palabras en inglés.

2. El hecho de que una sola palabra tenga más de una definición en muchos casos, nos hace ver lo complicado que puede llegar a ser el crear un traductor automático.

3. Como trabajo futuro, tener a mano todo el diccionario en ficheros de texto puede ser de gran ayuda a la hora de crear otro diccionario en otro idioma distinto al inglés.

5. CAPÍTULO

Clasificación de dominios sin entrenamiento

5.1. Introducción

En este capítulo trabajaremos con la clasificación de dominios del diccionario Oxford.[1] En el ámbito en el que trabajamos, un dominio indica la pertenencia a una categoría determinada de una palabra. Aquí veremos cuántos dominios contiene el diccionario, cuáles son los más frecuentes y cuáles coinciden con la clasificación creada por el programa Ask2Transformers.

El programa **Ask2Transformers** [Sainz and Rigau, 2020] - clasificador de dominios Zero-Shot con modelos de lenguaje pre-entrenados - desarrollado por Oscar Sainz y German Rigau anota automáticamente los datos textuales sin ninguna supervisión. Dado un conjunto particular de etiquetas (BabelDomains, WNDomains...), el sistema tiene que clasificar los datos sin ejemplos previos. Este trabajo utiliza la biblioteca Transformers y sus LM pre-entrenados. Evalúa los sistemas en el conjunto de datos BabelDomains logrando una precisión del 92,14 % en el etiquetado de dominios.[2]

El código generado por nosotros ha sido desarrollado en **Python** haciendo uso de la herramienta **Google Colaboratory**.[3] El código está subido en la plataforma de **GitHub**[4] y tanto el código como los resultados obtenidos han sido guardados en **Google Drive**. [5]

[1] https://www.lexico.com/
[2] https://github.com/osainz59/Ask2Transformers
[3] https://colab.research.google.com/
[4] https://github.com/leirevaran/lexico
[5] https://drive.google.com/drive/folders/1T-dBDwEtl_IVs0OytPycFedGj07YahIF?usp=

5.2. Desarrollo

En esta sección hablaremos sobre la clasificación de dominios que contiene el diccionario de Oxford. Explicaremos cómo los hemos obtenido, cuántos hay y con qué frecuencia aparecen en el diccionario. También hablaremos sobre el funcionamiento del clasificador de dominios Ask2Transformers [Sainz and Rigau, 2020] y comentaremos cómo lo hemos aplicado y evaluado nosotros.

5.2.1. Dominios Oxford

Como hemos visto en el capítulo anterior, partimos de la base de que ya tenemos unos ficheros de texto donde están almacenadas todas las definiciones de las palabras que contiene el diccionario de Oxford. Algunas de estas definiciones tienen asignados uno o varios dominios, los cuales también están anotados en estos mismos ficheros.

```
WORD: family    POS: noun       DOM: Biology
```

Figura 5.1: Dominio de la palabra *family* en el fichero

Por tanto, nuestro primer quehacer va a ser obtener todos los dominios del diccionario.

Para ello, hemos implementado un programa muy sencillo que se encarga de recorrer los ficheros correspondientes a cada letra y a buscar e imprimir en otro fichero de texto los dominios que encuentre.

Así ha sido la implementación:

- El programa recibe como parámetro el fichero del que vamos a sustraer los dominios. Va leyéndolo línea a línea buscando si hay algún dominio. Distinguiremos los dominios del resto de palabras porque siempre van seguidos de la nomenclatura *DOM:* al final de la línea (ver figura 5.1).

- En caso de que la línea contenga la nomenclatura *DOM:*, el programa imprimirá en el fichero de salida las palabras que haya a continuación, que serán las que se refieran a uno o varios dominios.

`sharing`

Una vez ejecutado este programa con los ficheros correspondientes a todas las letras del abecedario como entrada, se han obtenido otros veintisiete ficheros (uno por letra más los caracteres especiales) con todos los dominios existentes en el diccionario en línea.

Como era de esperar, muchos dominios se repetían en numerosas ocasiones, por lo que hemos generado un fichero de texto conteniendo todos los dominios del diccionario en orden y sin repetidos. Lo hemos generado con las siguientes instrucciones:

```
sort -gr    all_domains.txt                  >  domains1.txt
uniq -c     domains1.txt                     >  domains2.txt
awk    '{ print $2,$3,$4,$5 }' domains2.txt  >  domains3.txt
```

Vamos a suponer que *all_domains.txt* es el nombre del fichero que hemos obtenido juntando los resultados de los veintisiete ficheros. La primera instrucción ordena los dominios de *all_domains.txt* alfabéticamente pero en orden inverso. El resultado lo devuelve en *domains1.txt*.
La segunda instrucción cuenta y anota en *domains2.txt* cuántas veces aparece la misma palabra y elimina las líneas repetidas de *domains1.txt*.

```
1 Zoology Sociology
1 Zoology Psychology
7 Zoology Medicine
1 Zoology Entomology Botany
2 Zoology Entomology
```

Figura 5.2: Parte del fichero *domains2.txt*

El fichero *domains2.txt* contiene una primera columna con números (creada en la segunda instrucción) y en el resto de la línea se encuentran uno o varios dominios que corresponden a alguna palabra. La última instrucción se queda solo con las columnas de la 2 a la 5, es decir, las que contienen los dominios. Dicho de otra forma, elimina los números de la primera columna. El resultado final queda guardado en *domains3.txt*.

Dominios multipalabra

Hay ciertos dominios que están formados por dos o más palabras, como por ejemplo: *American Football*. Como algunas palabras contienen más de un dominio, y estos son separados por espacios en blanco, hemos revisado todos los dominios del fichero final (en el que salen en orden alfabético y sin repetidos) y sustituido los espacios en blanco que

había entre las palabras que conforman el dominio multipalabra por una barra baja. Por ejemplo:

American Football	→	American_Football
Roman Catholic Church Christianity	→	Roman_Catholic_Church Christianity

Lo siguiente que hemos hecho ha sido generar un código que se encargue de localizar los dominios multipalabra tanto en los ficheros que contienen el diccionario como en los que contienen todos los dominios de cada letra, y añadirles una barra baja entre medias.

La siguiente tabla muestra la cantidad de dominios que contiene el diccionario. Llamamos *dominios simples* a los dominios formados por una sola palabra.

	Cantidad dominios
Dominios simples	161
Dominios multipalabra	34
Total	195

Tabla 5.1: Cantidad dominios de Oxford

A continuación se muestran los 161 dominios simples que contiene el diccionario de Oxford junto con la cantidad de veces que aparecen en definiciones de palabras.

Dominio	Frecuencia	Dominio	Frecuencia
Accounting	18	Dance	10
Aeronautics	161	Darts	3
Alchemy	11	Dentistry	111
Anatomy	1867	Ecology	327
Anthropology	269	Economics	262
Archaeology	404	Electronics	434
Archery	17	Embryology	125
Architecture	371	Engineering	248
Art	160	Entomology	625
Astrology	23	Falconry	30
Astronomy	112	Farming	124
Audio	30	Fashion	19
Ballet	36	Fencing	36
Baseball	257	Finance	271
Basketball	88	Fishing	131
Billiards	36	Forestry	29
Biochemistry	1307	Freemasonry	5
Biology	2652	Games	13
Botany	2154	Gaming	16
Bowls	4	Genetics	284
Boxing	48	Geography	134
Bridge	112	Geology	1602
Buddhism	40	Geometry	195
Building	127	Golf	114
Bullfighting	24	Grammar	589
Business	84	Gymnastics	1
Cards	54	Heraldry	242
Chemistry	3592	Hinduism	83
Chess	100	History	96
Christianity	10	Hockey	71
Cinema	127	Horticulture	60
Climbing	42	Hunting	43
Clockmaking	28	Ichthyology	41
Clothing	7	Islam	53
Computing	1416	Jainism	1
Cooking	96	Journalism	10
Cricket	328	Judaism	95
Croquet	8	Knitting	12
Crystallography	96	Law	1460
Curling	4	Linguistics	683
Cycling	27	Literature	49

Tabla 5.2: Dominios simples de Oxford y la frecuencia con la que aparecen en él (1/2)

Dominio	Frecuencia	Dominio	Frecuencia
Logic	234	Riding	35
Marketing	15	Rowing	15
Mathematics	1057	Rugby	111
Mechanics	128	Sailing	107
Medicine	4688	Scouting	2
Metallurgy	191	Sculpture	7
Meteorology	129	Shinto	2
Microbiology	143	Shipbuilding	37
Military	532	Skating	11
Mineralogy	758	Skiing	28
Mining	233	Snooker	23
Motorsports	23	Soccer	119
Mountaineering	35	Sociology	132
Music	1057	Sport	212
Mycology	142	Squash	3
Mythology	17	Statistics	214
Nautical	555	Sumo	9
Needlework	32	Surfing	54
Oceanography	26	Surgery	301
Ophthalmology	78	Surveying	66
Optics	96	Swimming	9
Ornithology	159	Technology	20
Palmistry	8	Telecommunications	110
Parapsychology	33	Television	64
Pathology	169	Tennis	78
Pharmacology	345	Textiles	92
Philately	13	Theatre	65
Philosophy	729	Theology	62
Phonetics	285	Typography	41
Photography	246	University	22
Physics	1660	Veterinary	16
Physiology	827	Watchmaking	20
Politics	296	Weightlifting	14
Printing	270	Winemaking	11
Prosody	158	Woodworking	31
Psychiatry	128	Wrestling	20
Psychoanalysis	56	Yoga	7
Psychology	561	Zoology	3864
Radio	18		
Railways	31		
Rhetoric	69		

Tabla 5.3: Dominios simples de Oxford y la frecuencia con la que aparecen en él (2/2)

La siguiente tabla contiene los 34 dominios multipalabra que contiene el diccionario en línea y la cantidad de veces que han sido asignados a una definición.

Dominio	Frecuencia	Dominio	Frecuencia
Alternative_Medicine	11	Hat_Making	7
American_Football	205	Horse_Racing	45
Ancient_Greek_History	105	Ice_Hockey	52
Australian_Rules	36	Irish_Mythology	4
Bee_Keeping	11	Martial_Arts	23
Bell_Ringing	20	Political_Economy	14
British_Law	9	Roman_Catholic_Church	82
Canadian_Football	2	Roman_History	125
Carriage_Building	2	Roman_Mythology	43
Cheese_Making	5	Scandinavian_Mythology	19
Christian_Church	465	Science_Fiction	89
Christian_Science	1	Scots_Law	102
Christian_Theology	52	Sheep_Shearing	2
Egyptian_Mythology	23	Social_Sciences	30
Electrical_Engineering	57	Soil_Science	69
English_Law	52	Stock_Market	145
Greek_Mythology	273	Us_Law	39

Tabla 5.4: Dominios multipalabra de Oxford y la frecuencia con la que aparecen en él

El dominio que aparece con más frecuencia es *Medicine* (4688 apariciones). Los siguientes más frecuentes son *Zoology* (3864) y *Chemistry* (3592).
Los menos frecuentes son *Gymnastics*, *Jainism* (que es una religión de la India) y *Christian_Science*, que aparecen tan solo una vez.

5.2.2. Ask2Transformers

Dado un fichero con dominios y otro con palabras y sus respectivas definiciones, el programa Ask2Transformers se encarga de hacer una predicción que indica qué dominio tiene más relación con cada definición.

Por ejemplo, disponiendo de los siguientes dominios:

Dominio	Dominio
Administration	Literature
Animals	Mathematics
Anthropology	Medicine
Architecture	Military
Art	Mythology
Astronomy	Paranormal
Biology	Pedagogy
Chemistry	Philosophy
Commerce	Physics
Computer Science	Plants
Earth	Politics
Economy	Psychology
Engineering	Publishing
Environment	Radio
Farming	Religion
Fashion	Sexuality
Food	Sociology
Game	Sport
Health	Statistics
History	Telecommunication
Home	Theology
Industry	Tourism
Law	Transport
Linguistics	Tv

Tabla 5.5: Dominios de ejemplo

Y disponiendo de la siguiente definición:

- Hospital: a health facility where patients receive treatment.

Este sería el resultado que nos devolvería el programa:

```
hospital: a health facility where patients receive treatment.
[(0.63619065, 'health'),
 (0.32835743, 'medicine'),
 (0.0044441437, 'biology'),
 (0.0016410802, 'earth'),
 (0.0013727105, 'administration'),
 (0.001172708, 'anthropology'),
 (0.0011138729, 'history'),
 (0.0011125967, 'sociology'),
 (0.0011040716, 'linguistics'),
 (0.0010129844, 'literature'),
 (0.0009485316, 'statistics'),
 (0.00094460434, 'philosophy'),
 (0.00093834405, 'publishing'),
 (0.0008449697, 'mythology'),
 (0.00077979406, 'theology'),
 (0.00074205274, 'pedagogy'),
 (0.0007253911, 'radio'),
 (0.000724015, 'environment'),
 (0.00068564154, 'industry'),
 (0.0006664652, 'sexuality'),
 (0.0006444997, 'architecture'),
 (0.00059523596, 'home'),
 (0.00059492275, 'paranormal'),
 (0.0005901638, 'economy'),
 (0.00058689475, 'law'),
 (0.0005707025, 'tv'),
 (0.0005643218, 'telecommunication'),
 (0.0005633673, 'farming'),
 (0.00055981596, 'military'),
 (0.0005428184, 'psychology'),
 (0.00054082903, 'mathematics'),
 (0.00053997507, 'game'),
 (0.000531054, 'tourism'),
 (0.0005253325, 'commerce'),
 (0.0005063664, 'plants'),
 (0.000503849, 'transport'),
 (0.00050007086, 'engineering'),
 (0.0004940782, 'chemistry'),
 (0.00049259386, 'physics'),
 (0.00046857528, 'fashion'),
 (0.00046385702, 'astronomy'),
 (0.0004630824, 'religion'),
 (0.00045017578, 'art'),
 (0.00045015372, 'animals'),
 (0.00044879594, 'sport'),
 (0.0004375275, 'politics'),
 (0.00042479092, 'computer science'),
 (0.00042406234, 'food')]
```

Figura 5.3: Resultado de ejemplo de Ask2Transformers

Asignándole a la definición de *hospital* el dominio *health* como primero, y *food* como último.

Es importante remarcar dos características del fichero que contiene las palabras y las definiciones: una es que la si la entrada está formada por más de una palabra (por ejemplo: *flame out*), hay que rellenar los espacios en blanco que las separan con el símbolo '_'

(*flame_out*). La segunda es que en este fichero tampoco puede haber líneas en blanco, cada línea debe contener la entrada seguida de su definición.

Hemos hecho una ligera modificación en el programa cambiando *argv[1]* y *argv[2]* por el nombre del fichero de texto que contiene los dominios y el que contiene las palabras con sus definiciones. También le hemos añadido un tercer fichero de escritura que usaremos para imprimir el resultado.

Ahora que ya sabemos cómo funciona el programa Ask2Transformers, nuestro deber es probarlo con dominios y definiciones del diccionario de Oxford y ver si los resultados coinciden. Para ello, hemos ido seleccionando y anotando en un fichero definiciones que tuvieran ya asignado algún dominio, y en otro hemos anotado cada palabra seleccionada y el dominio que tuviera asignado.

```
cebid: A primate of a family (Cebidae ) that includes most of the New World monkeys.
family: A group of one or more parents and their children living together as a unit.
passback: A backward pass from the centre to put the ball in play.
```

```
cebid: Zoology
family: Biology
passback: American_football
```

Después, con el fichero de las definiciones y un listado de dominios, hemos puesto en marcha Ask2Transformers y hemos obtenido sus clasificaciones. En las siguientes imágenes se muestra la primera parte de cada resultado.

```
cebid: A primate of a family (Cebidae ) that includes most of the New World monkeys.
[(0.07554776, 'Zoology'),
```

```
family: A group of one or more parents and their children living together as a unit.
[(0.027557436, 'Biology'),
```

```
passback: A backward pass from the centre to put the ball in play.
[(0.03560896, 'Anatomy'),
```

Por último, debemos comparar las primeras clasificaciones de Ask2Transformers con los dominios asignados por Oxford. Para ello, hemos implementado un programa al que le pasamos toda la información en ficheros de texto y compara los resultados. En caso de que los dominios coincidan, devuelve *OK*. En caso contrario, *FAIL*.

```
cebid: Zoology, 0.07554776 Zoology  OK
family: Biology, 0.027557436 Biology  OK
passback: American_football, 0.03560896 Anatomy  FAIL
```

Figura 5.4: Resultado de comparar los dominios de Oxford con los de Ask2Transformers

5.3. Resultados

Hemos realizado distintas pruebas usando el programa Ask2Transformers y siguiendo los pasos descritos en el apartado 5.2.2. Para empezar, hemos hecho una pequeña prueba pasándole como entrada un fichero con diez definiciones que ya tuvieran asignado algún dominio en Oxford, y otro fichero con todos los dominios del mismo diccionario (195 en total).

Evidentemente, con una variedad de dominios tan grande el programa ha tardado mucho en clasificarlos todos con cada definición. Además, al ser algunos muy similares entre sí (por ejemplo: *Sailing* y *Nautical*), la primera clasificación de Ask2Transformers solo ha coincidido con la de Oxford en una ocasión.

Por tanto, en las siguientes pruebas hemos utilizado ficheros que contuvieran un número menor de dominios tanto para maximizar la probabilidad de que los dominios de Oxford coincidan con la clasificación de Ask2Transformers, como para ahorrar tiempo de ejecución.

5.3.1. Dominios en mayúsculas o en minúsculas

La primera prueba consiste en comprobar si existe alguna diferencia entre pasarle los dominios en mayúsculas o en minúsculas al programa Ask2Transformers. Para ello hemos creado dos ficheros con los mismos dominios, solo que en uno todos empiezan por letra mayúscula y en el otro por minúscula. Los dominios utilizados en esta prueba han sido los de la tabla 5.6. Hemos probado la misma entrada (*hospital*) con ambos ficheros y estos han sido los primeros veinte dominios que se han obtenido como resultado en cada caso:

```
hospital: a health facility where patients receive treatment.
[(0.45093137, 'Medicine'),
 (0.020600792, 'Physiology'),
 (0.014444758, 'Biology'),
 (0.011675493, 'Typography'),
 (0.010772647, 'Pharmacology'),
 (0.010406776, 'Logic'),
 (0.009947215, 'Anatomy'),
 (0.009270811, 'Phonetics'),
 (0.009122025, 'Grammar'),
 (0.008718171, 'Roman_Catholic_Church'),
 (0.008636255, 'Anthropology'),
 (0.008317719, 'Psychoanalysis'),
 (0.00820466, 'Australian_Rules'),
 (0.008181412, 'Geometry'),
 (0.007946548, 'Heraldry'),
 (0.007938485, 'Greek_Mythology'),
 (0.007864128, 'Building'),
 (0.007855784, 'Egyptian_Mythology'),
 (0.007445865, 'Science_Fiction'),
 (0.007398719, 'Palmistry'),
```

Figura 5.5: Resultado de *hospital* con dominios en mayúsculas

```
hospital: a health facility where patients receive treatment.
[(0.805985, 'medicine'),
 (0.032750405, 'physiology'),
 (0.010957695, 'biology'),
 (0.005408965, 'anatomy'),
 (0.0051830653, 'logic'),
 (0.0039569777, 'pharmacology'),
 (0.0033809063, 'typography'),
 (0.0031170675, 'anthropology'),
 (0.003060951, 'grammar'),
 (0.0029878833, 'building'),
 (0.0027982772, 'history'),
 (0.0027020462, 'linguistics'),
 (0.0026314396, 'roman_catholic_church'),
 (0.0025208092, 'phonetics'),
 (0.0024698537, 'philosophy'),
 (0.0024167777, 'geometry'),
 (0.0023981773, 'statistics'),
 (0.002205204, 'science_fiction'),
 (0.0022039404, 'crystallography'),
 (0.002099666, 'psychoanalysis'),
```

Figura 5.6: Resultado de *hospital* con dominios en minúsculas

Como podemos ver, los resultados cambian según cómo esté representado el dominio. Sí que es cierto que, los primeros tres resultados, en este caso: *Medicine*, *Physiology* y *Biology*, son iguales en ambos casos. Empiezan a ser distintos a partir del cuarto. Sin embargo, a pesar de ser iguales los tres primeros, su número de probabilidad cambia.

Para hacer el resto de pruebas, hemos usado siempre dominios en mayúsculas.

5.3.2. 100 definiciones y 90 dominios

Para evitar los problemas que comentábamos de que los dominios devueltos por el programa casi no coinciden con los de Oxford y que el programa tarda mucho en ejecutarse, hemos reducido los dominios a 90. Luego hemos seleccionado 100 definiciones del diccionario que tuvieran algún dominio asignado y las hemos dividido en dos ficheros de 50 que los pasaremos por separado a Ask2Transformers como parámetro. En otro fichero hemos anotado los dominios asignados por Oxford para poder comparárlos más tarde.

A continuación mostramos las 100 definiciones utilizadas y, más adelante, los dominios (tabla 5.6).

1. **earing**: In earlier use: a loop in the corner of a sail through which a rope may be attached. Later: any of various small ropes threaded through such a loop in order to fasten the (upper) corner of a sail to the yard.
2. **mentalism**: The theory that physical and psychological phenomena are ultimately explicable only in terms of a creative and interpretative mind.
3. **otalgia**: Earache.
4. **cebid**: A primate of a family (Cebidae) that includes most of the New World monkeys.
5. **dictionary**: A set of words or other text strings made for use in applications such as spellcheckers.
6. **passback**: A backward pass from the centre to put the ball in play.
7. **passion_day**: Originally: a day on which a Christian martyr suffered death, or on which the martyr's sufferings are commemorated. Later: (in extended use) of non-Christian martyrs.

8. **posthumanism**: The idea that humanity can be transformed, transcended, or eliminated either by technological advances or the evolutionary process; artistic, scientific, or philosophical practice which reflects this belief.

9. **family**: A group of one or more parents and their children living together as a unit.

10. **softballer**: A person who plays softball; a softball player.

11. **house**: A twelfth division of the celestial sphere, based on the positions of the ascendant and midheaven at a given time and place, and determined by any of a number of methods.

12. **housebote**: firebote.

13. **household_division**: A division comprising troops with (at least nominal) responsibility for guarding the monarch or head of state; specifically (in the British Army) the Household Cavalry, the Foot Guards, and (since 2004) the London Regiment of the Territorial Army.

14. **houseline**: A light line or rope of three strands, used for lashing, seizings, etc.; also called housing.

15. **wacke**: A sandstone of which the mud matrix in which the grains are embedded amounts to between 15 and 75 per cent of the mass.

16. **wage**: A fixed regular payment earned for work or services, typically paid on a daily or weekly basis.

17. **waggler**: A type of long float designed to be especially sensitive to movement of the bait, chiefly used in semi-still water.

18. **WAIS**: Wide area information service, designed to provide access to information across a computer network.

19. **xanthan_gum**: A substance produced by bacterial fermentation or synthetically and used in foods as a gelling agent and thickener. It is a polysaccharide composed of glucose, mannose, and glucuronic acid.

20. **xanthine**: A crystalline compound found in blood and urine which is an intermediate in the metabolic breakdown of nucleic acids to uric acid.

21. **naat**: An irregularity in the structure of a diamond, caused by a change of direction in the grain as a result of twinning.

22. **nabam**: A water-soluble powder which is a fungicide formerly sprayed on the leaves of plants but now applied to soil to protect against certain root rots; disodium ethylenebisdithiocarbamate, $(NaS{\cdot}CS{\cdot}NH{\cdot}CH_2 —)_2$.

23. **nabilone**: A synthetic cannabinoid, $C_{24}\ H_{36}\ O_3$, with anti-emetic properties, given orally for the relief of nausea and vomiting caused by cancer chemotherapy.

24. **nabism**: In form Nabism. The artistic theories or style of the Nabis.

25. **onagraceous**: Of or relating to the family Onagraceae, of which Onagra (now Oenothera) was the type genus, and which includes plants such as the evening primroses, willowherbs, and fuchsias.

26. **on-baller**: A player who follows the ball, rather than playing in a set position.

27. **on-beat**: Corresponding in time with the most accented beat in the bar.

28. **on-chip**: Denoting or relating to circuitry included in a single integrated circuit or in the same integrated circuit as a given device.

29. **ornithorhynchous**: Shaped like a bird's beak; having a beak like that of a bird.

30. **orocline**: An orogenic belt that has been curved or sharply bent.

31. **sense-appearance**: That which is perceived by or appears to the senses; also as a count noun.

32. **sense-feeling**: The feeling produced by one or more senses.

33. **solifluction**: The gradual movement of wet soil or other material down a slope, especially where frozen subsoil acts as a barrier to the percolation of water.

34. **solifuge**: A sun spider.

35. **stannic**: Of tin with a valency of four; of tin(IV).

36. **QSS**: Quasi-stellar source (of radio waves).

37. **quadded**: Especially of a cable: that is in the form of a quad; made as a quad.

38. **quadra**: A square border or frame round a bas-relief, panel, etc.; (more generally) a border or frame of any form.

39. **quare_impedit**: A writ issued in cases of disputed presentation to a benefice, requiring the defendant to state why he or she hinders the plaintiff from making the presentation; an action brought in such a case.

40. **quark_star**: A hypothetical celestial object, intermediate in density between a neutron star and a black hole, whose matter is in the form of quarks unconfined within nucleons.

41. **R0**: A figure expressing the average number of cases of an infectious disease arising by transmission from a single infected individual, in a population that has not previously encountered the disease.

42. **Ra**: The sun god, the supreme Egyptian deity, worshipped as the creator of all life and typically portrayed with a falcon's head bearing the solar disc. From earliest times he was associated with the pharaoh.

43. **rabatment**: The action of rabatting or rotating a plane, figure, etc.; the result of this.

44. **Rabbanite**: A Jewish person who accepts the teachings of the rabbis; a rabbinist. Opposed to Karaite.

45. **rabbit_ball**: A baseball deemed to have a livelier bounce than is usual; a ball that can be easily hit for a long distance.

46. **tabes**: Emaciation.

47. **table_lathe**: A small lathe that is clamped to a table when in use.

48. **table_line**: The line running from beneath the little finger to the base of the index or middle finger, forming the upper boundary of the table.

49. **thought-body**: A supposed spiritual counterpart of a physical body.

50. **titling_font**: A font consisting of full-faced letters, usually capitals, and typically used for titles and headings.

51. **Aanderaa**: Designating an instrument for measuring, recording, and transmitting oceanographic data, typically the direction and speed of currents and the water temperature; especially in .Aanderaa current meter".

52. **abactinal**: Designating the part or surface of an echinoderm or other radiate animal that is furthest from the mouth (as the upper side of a starfish); relating to this part.

53. **ad_court**: advantage court

54. **add-back**: Adjustment of net income through addition or deduction of items not affecting working capital; an item thus added or deducted.

55. **added_money**: Money added by a racing association, etc., to the stakes; frequently attributive.

56. **babbitting**: The action or process of adding a lining of babbitt metal to a bearing or other part in order to reduce friction.

57. **Babesia**: A genus of protozoans of the group Apicomplexa, the members of which are parasites of invertebrates and vertebrates causing babesiosis in domestic and wild mammals (and occasionally in humans); (also babesia) a protozoan of this genus.

58. **babingtonite**: A rare mineral occurring as dark greenish-black striated crystals, typically in association with zeolites.

59. **babord**: The port side of a ship.

60. **bed_and_breakfast**: Sell (shares) and buy them back by agreement the next day.

61. **celebret**: A document, signed and sealed by a bishop or other ecclesiastical authority, giving a priest permission to say mass in a diocese other than his own.

62. **cellarhood**: (originally and chiefly Baseball). The state of being in the lowest (or occasionally a low) position in the overall rankings or standings of a league or other grouping.

63. **Cellnet**: A cellular radio network used mainly for communication by mobile phone.

64. **cellobiase**: An enzyme that catalyses the hydrolysis of cellobiose into glucose.

65. **cellulin**: A complex of chitin and glucan, found in granules in cells and at hyphal constrictions in certain aquatic oomycetes. Usually attributive, especially in çellulin granule", çellulin plug".

66. **daunorubicin**: A synthetic antibiotic that interferes with DNA synthesis and is used in the treatment of acute leukaemia and other cancers.

67. **dayan**: A religious judge, in particular one in a rabbinic court.

68. **day_haul**: The action or an act of hauling a net for fish during the day as opposed to at night.

69. **day-hole**: The entrance to a day drift.

70. **daylighted**: Illuminated by (or as if by) daylight; exposed to daylight.

71. **egg-butt**: Used attributively to designate a type of snaffle in which the connection between the mouthpiece and each side-ring is an egg-shaped joint.

72. **ego**: A person's sense of self-esteem or self-importance.

73. **egressive**: (of a speech sound) produced using the normal outward-flowing airstream.

74. **eicosapentaenoate**: eicosapentaenoic_acid

75. **Embioptera**: A small order of insects that comprises the web-spinners.

76. **falcate**: Curved like a sickle; hooked.

77. **falcial**: Of or relating to the falx of the cerebrum.

78. **fallaway**: A shot made while the shooter jumps or falls away from the basket.

79. **fallibilism**: The principle that propositions concerning empirical knowledge can be accepted even though they cannot be proved with certainty.

80. **frisket**: A thin metal frame keeping the paper in position during printing on a hand press.

81. **Gouraud**: In computer graphics: designating or relating to a method of modelling the illumination of objects in order to simulate the effects of light and shade across their surface. Especially in "Gouraud shading".

82. **governor_block**: A (typically adjustable) block of metal or wood forming part of a governor.

83. **GPMG**: General-purpose machine gun, a machine gun designed to be capable of performing a variety of different combat roles.

84. **Graafian_follicle**: A fluid-filled structure in the mammalian ovary within which an ovum develops prior to ovulation.

85. **gracilis**: A slender superficial muscle of the inner thigh.

86. **indeclinable**: (of a noun, pronoun, or adjective in a highly inflected language) having no inflections.

87. **indecomposable**: Unable to be expressed as a product of factors or otherwise decomposed into simpler elements.

88. **indefeasible**: Not subject to being lost, annulled, or overturned.

89. **indehiscent**: (of a pod or fruit) not splitting open to release the seeds when ripe.

90. **indented**: Divided or edged with a zigzag line.

91. **j'adoube**: A declaration by a player intending to adjust the placing of a chessman without making a move with it.

92. **Jagannatha**: The form of Krishna worshipped in Puri, Orissa, where in the annual festival his image is dragged through the streets on a heavy chariot; devotees are said formerly to have thrown themselves under its wheels.

93. **Janus**: An ancient Italian deity, guardian of doorways and gates and protector of the state in time of war. He is usually represented with two faces, so that he looks both forwards and backwards.

94. **Jason**: The son of the king of Iolcos in Thessaly, and leader of the Argonauts in the quest for the Golden Fleece.

95. **jato**: Jet-assisted take-off.

96. **pistilliform**: Having the form of a pestle.

97. **pistillody**: The transformation of floral organs (usually stamens) into carpels.

98. **pistolgraph**: A camera for taking instantaneous (short exposure) photographs, especially of moving objects; a photograph taken by such a camera. Also figurative (usually attributive).

99. **pistomesite**: A mineral that is a magnesian variety of siderite.

100. **post-final**: A consonant following a main final consonant in a word-final consonant cluster.

Y estos han sido los 90 dominios con las que las hemos clasificado:

Dominio	Dominio	Dominio
Accounting	Electronics	Mineralogy
Aeronautics	Engineering	Mining
American_Football	Entomology	Music
Anatomy	Finance	Mycology
Anthropology	Fishing	Nautical
Archaeology	Football	Oceanography
Architecture	Forestry	Palmistry
Art	Geology	Parapsychology
Astrology	Geometry	Pharmacology
Astronomy Science	Grammar	Philosophy
Australian_Rules	Greek_Mythology	Phonetics
Baseball	Heraldry	Photography
Basketball	Hinduism	Psychoanalysis
Biochemistry	History	Physics
Biology	Horse_Racing	Politics
Botany	Hunting	Printing
Building	Islam	Psychology
Business	Journalism	Physiology
Chemistry	Judaism	Riding
Chess	Law	Roman_Catholic_Church
Christian_Church	Linguistics	Roman_Mythology
Cinema	Logic	Science_Fiction
Clothing	Marketing	Sport
Computing	Mathematics	Statistics
Crystallography	Mechanics	Stock_Market
Dance	Medicine	Technology
Dentistry	Metallurgy	Telecommunications
Ecology	Meteorology	Tennis
Economics	Microbiology	Typography
Egyptian_Mythology	Military	Zoology

Tabla 5.6: Dominios 5.3.2

Tras hacer las ejecuciones, estos han sido los resultados de hacer las comparaciones de los dominios de Ask2Transformers con los de Oxford.

```
earing: Nautical, 0.13919428 Nautical  OK
mentalism: Philosophy, 0.3019753 Psychology  FAIL
otalgia: Medicine, 0.19995534 Archaeology  FAIL
cebid: Zoology, 0.07554776 Zoology  OK
dictionary: Computing, 0.078272186 Linguistics  FAIL
passback: American_football, 0.03560896 Anatomy  FAIL
passion_day: Christian_Church, 0.06652379 Christian_Church  OK
posthumanism: Science_fiction, 0.14289472 Technology  FAIL
family: Biology, 0.027557436 Biology  OK
softballer: Sport, 0.06971413 Sport  OK
house: Astrology, 0.19082063 Building  FAIL
housebote: Law, 0.20194353 Nautical  FAIL
household_division: Military, 0.2877646 Politics  FAIL
houseline: Nautical, 0.16994654 Building  FAIL
wacke: Geology, 0.05206278 Mineralogy  FAIL
wage: Economics, 0.07976357 Business  FAIL
waggler: Fishing, 0.19087656 Nautical  FAIL
WISE: Computing, 0.72720885 Computing  OK
xanthan_gum: Chemistry, 0.06896027 Biology  FAIL
xanthine: Biochemistry, 0.1252658 Crystallography  FAIL
naat: Crystallography, 0.14002621 Grammar  FAIL
nabam: Chemistry, 0.08314914 Chemistry  OK
nabilone: Pharmacology, 0.19182892 Medicine  FAIL
nabism: Art, 0.88473606 Art  OK
onagraceous: Botany, 0.1756523 Botany  OK
on-baller: Australian_Rules Football, 0.027098903 Sport  FAIL
on-beat: Music, 0.048143566 Music  OK
on-chip: Electronics, 0.056482542 Electronics  OK
ornithorhynchous: Botany Zoology, 0.114559546 Anatomy  FAIL
orocline: Geology, 0.19457352 Geometry  FAIL
sense-appearance: Philosophy, 0.03430135 Physiology  FAIL
sense-feeling: Psychology, 0.10965199 Physiology  FAIL
solifluction: Geology, 0.058547318 Geometry  FAIL
solifuge: Zoology, 0.072791524 Entomology  FAIL
stannic: Chemistry, 0.054080524 Mineralogy  FAIL
QSS: Astronomy, 0.16083865 Astronomy  OK
quadded: Telecommunications, 0.04015162 Parapsychology  FAIL
quadra: Architecture, 0.32762146 Geometry  FAIL
quare_impedit: Law, 0.087106176 Law  OK
quark_star: Astrology, 0.066286415 Astronomy  FAIL
R0: Medicine, 0.07701981 Statistics  FAIL
Ra: Egyptian_Mythology, 0.3105557 Egyptian_Mythology  OK
rabatment: Geometry, 0.27742687 Geometry  OK
Rabbanite: Judaism, 0.6438329 Judaism  OK
rabbit_ball: Baseball, 0.74903613 Baseball  OK
tabes: Medicine, 0.08686299 Accounting  FAIL
table_lathe: Engineering, 0.042869 Parapsychology  FAIL
table_line: Palmistry, 0.07965996 Geometry  FAIL
thought-body: Parapsychology, 0.119020976 Parapsychology  OK
titling_font: Typography, 0.054506037 Typography  OK


38.0 % of domains are the same.
62.0 % of domains are different.
```

```
Aanderaa: Oceanography, 0.58282495 Oceanography  OK
abactinal: Zoology, 0.18195534 Nautical  FAIL
ad_court: Tennis, 0.073547974 Marketing  FAIL
add-back: Accounting, 0.06474837 Finance  FAIL
added_money: Horse_Racing, 0.5212217 Riding  FAIL
babbitting: Engineering, 0.03697546 Metallurgy  FAIL
Babesia: Microbiology, 0.13427204 Entomology  FAIL
babingtonite: Mineralogy, 0.045425255 Crystallography  FAIL
babord: Nautical, 0.58970445 Nautical  OK
bed_and_breakfast: Stock_Market, 0.0700375 Stock_Market  OK
celebret: Roman_Catholic_Church, 0.072007045 Heraldry  FAIL
cellarhood: Sport, 0.4573815 Baseball  FAIL
Cellnet: Telecommunications, 0.31399333 Telecommunications  OK
cellobiase: Biochemistry, 0.04504338 Biology  FAIL
cellulin: Mycology, 0.094567604 Nautical  FAIL
daunorubicin: Medicine, 0.27304396 Medicine  OK
dayan: Judaism, 0.42874107 Islam  FAIL
day_haul: Fishing, 0.34685633 Fishing  OK
day-hole: Mining, 0.4159157 Riding  FAIL
daylighted: Architecture, 0.03798511 Logic  FAIL
egg-butt: Riding, 0.09465738 Dentistry  FAIL
ego: Psychoanalysis, 0.028367557 Crystallography  FAIL
egressive: Phonetics, 0.04378981 Phonetics  OK
eicosapentaenoate: Chemistry, 0.048817337 Chemistry  OK
Embioptera: Entomology, 0.09781483 Entomology  OK
falcate: Botany Zoology, 0.32043287 Geometry  FAIL
falcial: Anatomy, 0.114128634 Biology  FAIL
fallaway: Basketball, 0.21298926 Photography  FAIL
fallibilism: Philosophy, 0.05435797 Logic  FAIL
frisket: Printing, 0.5216033 Printing  OK
Gouraud: Computing, 0.92471033 Computing  OK
governor_block: Mechanics, 0.035766855 Building  FAIL
GPMG: Military, 0.0714598 Military  OK
Graafian_follicle: Physiology, 0.040178027 Physiology  OK
gracilis: Anatomy, 0.087188445 Anatomy  OK
indeclinable: Grammar, 0.05193494 Phonetics  FAIL
indecomposable: Mathematics, 0.040399577 Roman_Catholic_Church  FAIL
indefeasible: Law Philosophy, 0.04068795 Law  OK
indehiscent: Botany, 0.07319425 Entomology  FAIL
indented: Heraldry, 0.12938762 Geometry  FAIL
j'adoube: Chess, 0.84731686 Chess  OK
Jagannatha: Hinduism, 0.64190227 Hinduism  OK
Janus: Roman_Mythology, 0.15818854 Roman_Mythology  OK
Jason: Greek_Mythology, 0.07687467 Military  FAIL
jato: Aeronautics, 0.29047233 Aeronautics  OK
pistilliform: Archaeology, 0.086558424 Anatomy  FAIL
pistillody: Botany, 0.25411066 Palmistry  FAIL
pistolgraph: Photography, 0.37052372 Photography  OK
pistomesite: Mineralogy, 0.040339254 Geology  FAIL
post-final: Linguistics, 0.031292804 Entomology  FAIL


40.0 % of domains are the same.
60.0 % of domains are different.
```

Es decir, en esta prueba de 100 definiciones y 90 dominios, en total coinciden los dominios de Oxford con los de Ask2Transformers un 39 % de las veces. Los resultados han mejorado en cuanto a la primera prueba que hemos hecho con los 195 dominios del diccionario y 10 definiciones, que solo coincidían 1/10. Sin embargo, todavía no alcanzamos

la mitad de acierto, por lo que o bien siguen siendo demasiados dominios, o todavía hay muchos que se parecen entre sí. Por ejemplo, la entrada *added_money* da *FAIL* por ser el dominio de Oxford *Horse_Racing* y el de Ask2Transformers *Riding*.

5.3.3. Agrupación de dominios

Con la intención de evitar que los dominios de Oxford y los de Ask2Transformers no coincidan debido a que son muy parecidos pero no iguales, y así aumentar la tasa de acierto, hemos reducido aún más el número de dominios agrupándolos como mostramos a continuación.

American_Football, Australian_Rules, Football, Tennis, Baseball, Basketball → **Sport**
Roman_Mythology, Egyptian_Mythology, Greek_Mythology → **Mythology**
Roman_Catholic_Church, Judaism, Islam, Hinduism, Christian_Church → **Religion**
Accounting, Business, Finance, Stock_Market, Marketing → **Economics**
Stock_Market → **Clothing**
Fishing, Oceanography → **Nautical**
Anatomy, Pharmacology, Physiology, Dentistry → **Medicine**
Archaeology, Forestry, Mineralogy, Mining, Crystallography → **Geology**
Entomology, Astronomy, Biology, Microbiology, Mycology, Science_Fiction → **Science**
Psychoanalysis, Parapsychology, Logic → **Psychology**
Grammar → **Linguistics**
Biochemistry → **Chemistry**
Heraldry → **History**
Horse_Racing → **Riding**

Después de estas agrupaciones, hemos reducido el número de dominios a la mitad y ahora son 45. Con esta nueva lista de dominios, hemos vuelto a probar las mismas 100 definiciones del apartado 5.3.2 para ver si coinciden más a menudo los dominios del diccionario con los del programa.

Estos han sido los resultados:

```
earing: Nautical, 0.34254047 Nautical  OK
mentalism: Philosophy, 0.36905563 Psychology  FAIL
otalgia: Medicine, 0.19326782 Astrology  FAIL
cebid: Zoology, 0.17742755 Zoology  OK
dictionary: Computing, 0.11654623 Linguistics  FAIL
passback: Sport, 0.06844258 Geometry  FAIL
passion_day: Religion, 0.65872246 Religion  OK
posthumanism: Science, 0.23230352 Technology  FAIL
family: Science, 0.03945357 Typography  FAIL
softballer: Sport, 0.20412968 Sport  OK
house: Astrology, 0.3047928 Building  FAIL
housebote: Law, 0.24589893 Nautical  FAIL
household_division: Military, 0.51358354 Military  OK
houseline: Nautical, 0.30562562 Building  FAIL
wacke: Geology, 0.09688732 Geology  OK
wage: Economics, 0.07120567 Economics  OK
waggler: Nautical, 0.44562373 Nautical  OK
WISE: Computing, 0.8434815 Computing  OK
xanthan_gum: Chemistry, 0.11643173 Chemistry  OK
xanthine: Chemistry, 0.07426104 Chemistry  OK
naat: Geology, 0.1979082 Geometry  FAIL
nabam: Chemistry, 0.22545277 Chemistry  OK
nabilone: Medicine, 0.40165174 Medicine  OK
nabism: Art, 0.9346425 Art  OK
onagraceous: Botany, 0.4160715 Botany  OK
on-baller: Sport, 0.0507631 Sport  OK
on-beat: Music, 0.10652679 Music  OK
on-chip: Electronics, 0.10780598 Electronics  OK
ornithorhynchous: Botany Zoology, 0.13786191 Geometry  FAIL
orocline: Geology, 0.25093353 Geometry  FAIL
sense-appearance: Philosophy, 0.058586203 Palmistry  FAIL
sense-feeling: Psychology, 0.07487588 Psychology  OK
solifluction: Geology, 0.13685851 Geometry  FAIL
solifuge: Zoology, 0.07214742 Meteorology  FAIL
stannic: Chemistry, 0.06710157 Chemistry  OK
QSS: Astronomy, 0.10372868 Chess  FAIL
quadded: Telecommunications, 0.078280784 Nautical  FAIL
quadra: Architecture, 0.5422079 Geometry  FAIL
quare_impedit: Law, 0.15093671 Law  OK
quark_star: Astrology, 0.08878887 Physics  FAIL
R0: Medicine, 0.05012262 Medicine  OK
Ra: Mythology, 0.17466214 Mythology  OK
rabatment: Geometry, 0.35410872 Geometry  OK
Rabbanite: Religion, 0.7360638 Religion  OK
rabbit_ball: Sport, 0.058352873 Nautical  FAIL
tabes: Medicine, 0.104463324 Ecology  FAIL
table_lathe: Engineering, 0.0749302 Geometry  FAIL
table_line: Palmistry, 0.16294216 Geometry  FAIL
thought-body: Psychology, 0.10110695 Psychology  OK
titling_font: Typography, 0.11537688 Typography  OK


54.0 % of domains are the same.
46.0 % of domains are different.
```

```
Aanderaa: Nautical, 0.9111769 Nautical  OK
abactinal: Zoology, 0.41206402 Nautical  FAIL
ad_court: Sport, 0.13948414 Law  FAIL
add-back: Economics, 0.08294653 Economics  OK
added_money: Economics Riding, 0.61789507 Riding  OK
babbitting: Engineering, 0.061232876 Nautical  FAIL
Babesia: Science, 0.20441528 Zoology  FAIL
babingtonite: Geology, 0.0620176 Nautical  FAIL
babord: Nautical, 0.6886276 Nautical  OK
bed_and_breakfast: Economics Clothing, 0.04477449 Building  FAIL
celebret: Religion, 0.10350424 Religion  OK
cellarhood: Sport, 0.674909 Sport  OK
Cellnet: Telecommunications, 0.35246095 Telecommunications  OK
cellobiase: Chemistry, 0.070777066 Chemistry  OK
cellulin: Science, 0.28212714 Nautical  FAIL
daunorubicin: Medicine, 0.5012748 Medicine  OK
dayan: Religion, 0.925776 Religion  OK
day_haul: Nautical, 0.30323386 Nautical  OK
day-hole: Geology, 0.5014793 Riding  FAIL
daylighted: Architecture, 0.06283264 Phonetics  FAIL
egg-butt: Riding, 0.06323333 Nautical  FAIL
ego: Psychology, 0.051498916 Psychology  OK
egressive: Phonetics, 0.08483997 Phonetics  OK
eicosapentaenoate: Chemistry, 0.10782288 Chemistry  OK
Embioptera: Science, 0.12738793 Nautical  FAIL
falcate: Botany Zoology, 0.34840444 Geometry  FAIL
falcial: Medicine, 0.09597046 Military  FAIL
fallaway: Sport, 0.29900247 Photography  FAIL
fallibilism: Philosophy, 0.07481413 Law  FAIL
frisket: Printing, 0.7030694 Printing  OK
Gouraud: Computing, 0.94526565 Computing  OK
governor_block: Mechanics, 0.05970711 Building  FAIL
GPMG: Military, 0.09966845 Military  OK
Graafian_follicle: Medicine, 0.076054476 Geometry  FAIL
gracilis: Medicine, 0.09278396 Geometry  FAIL
indeclinable: Linguistics, 0.10093038 Phonetics  FAIL
indecomposable: Mathematics, 0.07351091 Nautical  FAIL
indefeasible: Law Philosophy, 0.09458724 Law  OK
indehiscent: Botany, 0.12195885 Botany  OK
indented: History, 0.3001099 Geometry  FAIL
j'adoube: Chess, 0.91050625 Chess  OK
Jagannatha: Religion, 0.10298717 Religion  OK
Janus: Mythology, 0.13709301 Building  FAIL
Jason: Mythology, 0.181465 Military  FAIL
jato: Aeronautics, 0.381292 Aeronautics  OK
pistilliform: Geology, 0.13540314 Geometry  FAIL
pistillody: Botany, 0.23663023 Palmistry  FAIL
pistolgraph: Photography, 0.37840006 Photography  OK
pistomesite: Geology, 0.085758016 Geology  OK
post-final: Linguistics, 0.05160427 Palmistry  FAIL


48.0 % of domains are the same.
52.0 % of domains are different.
```

Como podemos ver, ha mejorado la tasa de acierto, ya que con la agrupación de dominios realizada y las mismas 100 definiciones utilizadas en el apartado anterior (5.3.2), ahora coinciden los dominios de Oxford con los de Ask2Transformers un 51 % de las veces. Al reducir el número de dominios a la mitad, hemos obtenido una tasa de acierto 12 % mayor

y los dominios coinciden más de la mitad de veces.

5.3.4. Dominios más frecuentes

Otra manera de evaluar la clasificación de dominios es seleccionando tan solo los dominios más frecuentes de Oxford. En esta prueba hemos seleccionado todos los dominios con una frecuencia mayor que 200 (ver tablas 5.2, 5.3 y 5.4). En total nos han quedado los siguientes 47 dominios.

Dominio	Dominio
Anatomy	Law
Anthropology	Linguistics
American_Football	Logic
Archaeology	Mathematics
Architecture	Medicine
Baseball	Military
Biochemistry	Mineralogy
Biology	Mining
Botany	Music
Chemistry	Nautical
Christian_Church	Pharmacology
Computing	Philosophy
Cricket	Phonetics
Ecology	Photography
Economics	Physics
Electronics	Physiology
Engineering	Politics
Entomology	Printing
Finance	Psychology
Genetics	Sport
Geology	Statistics
Grammar	Surgery
Greek_Mythology	Zoology
Heraldry	

Tabla 5.7: Dominios 5.3.4

Para la evaluación, hemos seleccionado otras 50 definiciones (distintas a las de las pruebas anteriores) que tuvieran asignado algún dominio de la tabla 5.7. Las hemos anotado en un fichero y las hemos probado es Ask2Transformers.

Estas son las 50 definiciones:

1. **all-Pro**: Designating, relating to, or consisting of the professional players considered to be the best in their positions, usually as determined by a vote of sportswriters, especially designating a (notional) team made of up such players.

2. **barrel_vault**: A vault forming a half cylinder.

3. **ferial**: Denoting an ordinary weekday, as opposed to one appointed for a festival or fast.

4. **halophile**: An organism, especially a microorganism, that grows in or can tolerate saline conditions.

5. **junk_DNA**: DNA that does not code for a protein, usually occurs in repetitive sequences of nucleotides, and does not seem to serve any useful purpose.

6. **King_of_Arms**: (in the UK) a chief herald. Those now at the College of Arms are the Garter, Clarenceux, and Norroy and Ulster Kings of Arms; the Lyon King of Arms has jurisdiction in Scotland.

7. **lentic**: (of organisms or habitats) inhabiting or situated in still fresh water.

8. **objectual**: Referring or relating to a material object, as opposed to a symbol or fictive referent.

9. **quartz_glass**: Glass composed of silica; silica glass.

10. **variate**: A quantity having a numerical value for each member of a group, especially one whose values occur according to a frequency distribution.

11. **deliberative_democracy**: A form of democracy in which authentic debate, usually aiming to reach a consensus, is used in decision-making; (sometimes) specifically a form of democracy in which citizens' assemblies form part of the decision-making process.

12. **extra-nidal**: Occurring or living outside the nest (of a social insect).

13. **corridor_of_uncertainty**: An area just outside the batsman's off stump, commonly used as a line of delivery by the bowler with the intention of leaving the batsman uncertain whether or not to play a shot; (also in extended use) a situation or course of action which causes hesitation or uncertainty over how to proceed.

14. **Ginnie_Mae**: The Government National Mortgage Association, a U.S. government agency established in 1968 as part of the restructuring of the Federal National Mortgage Association, which guarantees securities backed by government-insured mortgages and administers housing assistance programmes.

15. **intertrigo**: Inflammation caused by the rubbing of one area of skin on another.

16. **main_guard**: A body of troops constituting the chief guard, especially a body of cavalry posted on the wings of a camp towards the enemy; a guard in a fortress taking custody of disturbers of the peace, etc.

17. **nonfeasance**: Failure to perform an act that is required by law.

18. **Oamaru**: Designating a division of Tertiary rock formations found in the region of Oamaru; specifically (especially in .ºamaru stone") a white limestone quarried near Oamaru and formerly much used as a building material.

19. **pair_production**: The conversion of a radiation quantum into an electron and a positron.

20. **Wallacea**: A zoogeographical area constituting a transition zone between the Oriental and Australian regions, east of Wallace's line. It is generally held to comprise Sulawesi and other islands between the two continental shelves.

21. **telencephalon**: The most highly developed and anterior part of the forebrain, consisting chiefly of the cerebral hemispheres.

22. **obelion**: A point on the sagittal suture of the skull in line with the parietal foramina.

23. **racloir**: A scraper of a type discovered amongst the Mousterian remains of the Middle Palaeolithic period.

24. **metaplasm**: Usually with reference to classical languages: the alteration of a word by addition, removal, or transposition of letters or syllables; an instance of this.

25. **Gaia**: The Earth personified as a goddess, daughter of Chaos. She was the mother and wife of Uranus (Heaven); their offspring included the Titans and the Cyclops.

26. **kairomone**: A chemical substance emitted by an organism and detected by another of a different species which gains advantage from this, e.g. a parasite seeking a host.

27. **film_negative**: A photographic negative recorded on film.

28. **lah**: (in tonic sol-fa) the sixth note of a major scale.

29. **yo-hope**: A task regulated by a chant or song of the type used by sailors when hauling ropes or performing other strenuous, rhythmically repetitive tasks; the chant so used.

30. **vertical_market**: A market comprising all the potential purchasers in a particular occupation or industry.

31. **selectional**: Denoting or relating to the process by which only certain words or structures can occur naturally, normally, or correctly in the context of other words.

32. **payability**: Of a mine or mining area: the capacity to be worked profitably.

33. **Zephiran**: A preparation of benzalkonium chloride used as an antiseptic.

34. **acatalepsy**: Unknowability, incomprehensibility, originally as a characteristic of all things, according to the ancient Sceptics. Hence also: scepticism, profession of ignorance.

35. **idemfactor**: An operator or quantity which, when applied to another quantity, leaves that quantity unchanged.

36. **firmware**: Permanent software programmed into a read-only memory.

37. **limonene**: A colourless liquid hydrocarbon with a lemon-like scent, present in lemon oil, orange oil, and similar essential oils.

38. **quiteron**: A superconducting device with switching and amplifying characteristics similar to those of a transistor, but capable of operating at lower power levels.

39. **talik**: An area of unfrozen ground surrounded by permafrost.

40. **juncture**: The set of features in speech that enable a hearer to detect a word or phrase boundary (e.g. distinguishing I scream from ice cream).

41. **nervism**: A theory (advocated mainly by I. P. Pavlov and his followers) ascribing the control of most normal and abnormal physiological processes to the central nervous system.

42. **chemigrapher**: A person who produces printing plates by means of chemigraphy.

43. **allocentric**: Concentrating on or interested in external objects in themselves, rather than in regard to their relation or relevance to oneself.

44. **enterocele**: Herniation or prolapse of a segment of intestine; (in later use) especially prolapse of intestine into the rectouterine pouch; an instance or case of this.

45. **linespeople**: (originally and chiefly Tennis). In games played on a field or court: officials who assist the referee or umpire from the sideline or touchline, considered collectively.

46. **kanamycin**: A broad-spectrum antibiotic obtained from a strain of bacteria.

47. **rachiglossate**: Designating a radula bearing a median series of single teeth, often flanked by one row of lateral teeth on each side; having such a radula, typical of most neogastropods.

48. **rachilla**: In grasses: the floral axis, i.e. the axis on which the florets are borne within the spikelet. Also (in palms): the ultimate flower-bearing axis of the inflorescence.

49. **semantron**: A wooden or metal bar which produces a sound when struck by a mallet, and is used to summon worshippers to service.

50. **mascaron**: A representation, often grotesque, of a human or animal face, used as an architectural ornament.

Tras hacer las ejecuciones, estos han sido los resultados de hacer las comparaciones de los dominios de Ask2Transformers con los de Oxford.

```
all-Pro: American_Football, 0.78574854 Sport  FAIL
barrel_vault: Architecture, 0.07252116 Nautical  FAIL
ferial: Christian_Church, 0.050094794 Entomology  FAIL
halophile: Ecology, 0.09192484 Ecology  OK
junk_DNA: Genetics, 0.078209646 Genetics  OK
King_of_Arms: Heraldry, 0.8161704 Heraldry  OK
lentic: Ecology, 0.1453876 Nautical  FAIL
objectual: Logic, 0.049475808 Christian_Church  FAIL
quartz_glass: Mineralogy, 0.13222726 Chemistry  FAIL
variate: Statistics, 0.083225474 Statistics  OK
deliberative_democracy: Politics, 0.117285356 Politics  OK
extra-nidal: Entomology Ecology, 0.073082626 Nautical  FAIL
corridor_of_uncertainty: Cricket, 0.050375767 Cricket  OK
Ginnie_Mae: Finance, 0.15575604 Finance  OK
intertrigo: Medicine, 0.1179939 Physiology  FAIL
main_guard: Military, 0.6584228 Military  OK
nonfeasance: Law, 0.9266529 Law  OK
Oamaru: Geology Building, 0.47309574 Mineralogy  FAIL
pair_production: Physics, 0.09917594 Physics  OK
Wallacea: Zoology, 0.85580504 Zoology  OK
telencephalon: Anatomy, 0.07752659 Anatomy  OK
obelion: Anthropology Anatomy, 0.120611735 Anatomy  OK
racloir: Archaeology, 0.116104774 Archaeology  OK
metaplasm: Grammar, 0.5170722 Linguistics  FAIL
Gaia: Greek_Mythology, 0.14424014 Greek_Mythology  OK
kairomone: Biology, 0.13817954 Genetics  FAIL
film_negative: Photography, 0.70609444 Photography  OK
lah: Music, 0.074677065 Phonetics  FAIL
yo-hope: Nautical, 0.8405976 Nautical  OK
vertical_market: Economics, 0.12787765 Economics  OK
selectional: Linguistics, 0.49493033 Linguistics  OK
payability: Mining, 0.8012015 Mining  OK
Zephiran: Pharmacology, 0.24665508 Medicine  FAIL
acatalepsy: Philosophy, 0.1251903 Philosophy  OK
idemfactor: Mathematics, 0.03944581 Logic  FAIL
firmware: Computing, 0.1453033 Computing  OK
limonene: Chemistry, 0.054197878 Chemistry  OK
quiteron: Electronics, 0.09092254 Electronics  OK
talik: Geology, 0.09959203 Geology  OK
juncture: Phonetics, 0.30986956 Phonetics  OK
nervism: Physiology, 0.5684429 Physiology  OK
chemigrapher: Printing, 0.6037467 Printing  OK
allocentric: Psychology, 0.07626184 Logic  FAIL
enterocele: Surgery, 0.09607389 Anatomy  FAIL
linespeople: Sport, 0.73239046 Sport  OK
kanamycin: Medicine, 0.4361134 Biology  FAIL
rachiglossate: Zoology, 0.10063899 Nautical  FAIL
rachilla: Botany, 0.19919303 Entomology  FAIL
semantron: Christian_Church, 0.18041842 Nautical  FAIL
mascaron: Architecture, 0.85685056 Architecture  OK


60.0 % of domains are the same.
40.0 % of domains are different.
```

Esta vez hay un acierto del 60 %, lo cual significa que hemos obtenido el mejor resultado hasta ahora. Hemos obtenido una tasa de acierto 9 % y 21 % mayor en comparación con los apartados anteriores, 5.3.3 y 5.3.2 respectivamente.

Si nos fijamos en los fallos, vemos que de nuevo se repite el error de que entre los dominios más frecuentes también hay algunos muy parecidos entre sí. Esto pasa por ejemplo con la entrada *all-Pro*, que según Oxford pertenece al dominio *American_Football* y según la primera clasificación de Ask2Transformers, a *Sport*. Lo mismo pasa con las entradas *intertrigo* (*Medicine* y *Physiology*), *Oamaru* (*Geology* y *Mineralogy*), *metaplasm* (*Grammar* y *Linguistics*), *kairomone* (*Biology* y *Genetics*), *lah* (*Music* y *Phonetics*), *Zephiran* (*Pharmacology* y *Medicine*), *allocentric* (*Psychology* y *Logic*) y *enterocele* (*Surgery* y *Anatomy*). Si agrupásemos estos dominios como hemos hecho en el apartado 5.3.3, todas estas entradas darían *OK* y por tanto en esta prueba habría un acierto del 78 %.

5.3.5. Etiquetado de ejemplos

A veces puede que la definición sea muy corta o poco clara. En estos casos solamente con la definición el algoritmo no tiene suficiente información para elegir bien el dominio. Por lo tanto, esta última prueba consiste en probar a clasificar las frases de ejemplo aplicando el mismo procedimiento que hemos seguido hasta ahora con las definiciones. Lo único que va a cambiar es que obtendremos por cada palabra varios resultados dependiendo del número de ejemplos que usemos. Para agrupar estos resultados, nos fijaremos en la cantidad de veces que aparece un dominio como primero en la clasificación y en su peso. Por ejemplo, la entrada ***otalgia**: Earache* tiene las siguientes frases de ejemplo:

1. The main presenting symptoms were otalgia, sensation of blocked ears, hearing loss, otorrhoea and pruritis.
2. In our study, the main presenting symptoms were otalgia and sensation of blockage.
3. They were advised to use antibiotics if their child had severe otalgia or fever after 72 hours or if discharge lasted for 10 days or more.
4. Patients with external otitis complain of otalgia and sensitivity to auricular movement.
5. They studied children brought to family physicians in southern England with acute otalgia.

Si le pasamos estas frases a Ask2Transformers haciendo uso de los dominios más frecuentes (5.7) y comparamos los resultados con los de Oxford, obtenemos los siguientes resultados:

```
earache: Medicine, 0.06192145 Anatomy  FAIL
earache: Medicine, 0.06445137 Physiology  FAIL
earache: Medicine, 0.083124064 Medicine  OK
earache: Medicine, 0.05359797 Physiology  FAIL
earache: Medicine, 0.2766082 Medicine  OK

40.0 % of domains are the same.
60.0 % of domains are different.
```

Como podemos ver, todos los dominios asignados por Ask2Transformers, *Anatomy*, *Physiology* y *Medicine*, están relacionados con la medicina. Además, con un peso de 0.276608 y 0.08314064, las dos veces que el algoritmo ha etiquetado las frases como *Medicine* son las que mayor peso tienen. Por lo tanto, podríamos decir que los dominios asignados tanto por Oxford como por Ask2Transformers de la entrada *otalgia* coinciden en que ambos pertenecen a *Medicine*.

Esto supone una mejora respecto al etiquetado que se le había hecho a la definición de *otalgia*, que es simplemente la palabra *Earache*. Recordamos que cuando hemos probado esta definición en el apartado 5.3.3 de agrupación de dominios, el resultado que obteníamos era *Astrology*. Es decir, un dominio que no tiene nada que ver con el significado de la palabra. Lo mismo pasaba en la primera sección en la que usábamos 90 dominios 5.3.2, que el dominio asignado por Ask2Transformers era *Archaeology*.

Para las siguientes pruebas, vamos a generar una nueva agrupación de los dominios más frecuentes como hemos hecho en la sección 5.3.3.

American_Football, Baseball, Cricket → **Sport**
Anatomy, Pharmacology, Physiology, Surgery → **Medicine**
Biochemistry → **Chemistry**
Finance → **Economics**
Archaeology, Mining, Mineralogy → **Geology**
Music → **Phonetics**
Statistics → **Mathematics**
Logic → **Psychology**
Anthropology → **Biology**
Entomology → **Zoology**

Sin embargo, esta agrupación la consultaremos solo para ver el índice de error en la comparación de resultados, al igual que hemos hecho con la entrada *otalgia*. Los dominios que vamos a usar van a seguir siendo los 47 más frecuentes, representados en la tabla 5.7.

Las próximas entradas que vamos a probar van a ser los ejemplos de *egressive*.

***egressive**: (of a speech sound) produced using the normal outward-flowing airstream.*

1. Pulmonic egressive sounds are found in all human languages.
2. We examine how movement patterns are modified when speakers change from an egressive to ingressive airstream.

```
egressive: Phonetics, 0.08456217 Linguistics  FAIL
egressive: Phonetics, 0.24781427 Linguistics  FAIL

0.0 % of domains are the same.
100.0 % of domains are different.
```

Las clasificaciones no coinciden. De todos modos, la clasificación hecha por el programa tiene sentido, ya que en ambos ejemplos se mencionan palabras relacionadas con la lingüística como *languages* y *speakers*. En comparación con los resultados obtenidos de esta misma entrada usando su definición en apartados anteriores, el resultado ha empeorado. En estos casos anteriores la definición era clasificada con el dominio *Phonetics* coincidiendo así con el de Oxford.

El siguiente caso que vamos a analizar es especial, ya que la entrada no tiene una definición como tal, sino una referencia. Por lo tanto, hemos cogido los ejemplos de la palabra a la que hace referencia: *eicosapentaenoic_acid*.

***eicosapentaenoate**: eicosapentaenoic_acid*

1. Oils from deep sea fish are a rich source of omega - 3 polyunsaturated fatty acids, particularly eicosapentaenoic acid and docosahexaenoic acid.
2. Your brain needs long-chain polyunsaturated fatty acids to function properly and two types - eicosapentaenoic acid and docosahexaenoic acid - have come up trumps.
3. According to a new book, eicosapentaenoic acid is a substance that can ‘significantly alleviate the symptoms of depression, even in its most severe forms.’
4. Fish is the recommended source of omega - 3s because only marine life contains eicosapentaenoic acid and docosahexaenoic acid, the two most accessible forms of the fats.
5. Fish provide varying amounts of omega-3 fatty acids in the form of docosahexaenoic acid and eicosapentaenoic acid.

```
eicosapentaenoate: Chemistry, 0.121425815 Nautical  FAIL
eicosapentaenoate: Chemistry, 0.108273834 Physiology  FAIL
eicosapentaenoate: Chemistry, 0.08077602 Logic  FAIL
eicosapentaenoate: Chemistry, 0.15160811 Nautical  FAIL
eicosapentaenoate: Chemistry, 0.043835532 Physiology  FAIL

0.0 % of domains are the same.
100.0 % of domains are different.
```

El dominio que más peso tiene es *Nautical* y las clasificaciones no coinciden en ningún caso. Los ejemplos contienen cierto contenido que hace que el programa se centre en otros dominios. Por eso, en este caso es mejor usar la referencia, aun siendo esta una sola palabra. La salida *OK* obtenida en las pruebas anteriores al coincidir los dominios usando la referencia nos lo confirman.

***falcate**: Curved like a sickle; hooked.*

1. the mandibles are falcate
2. Beaked whales have a small, falcate dorsal fin, which is set fairly far back on their bodies (well beyond the midpoint).
3. The shape of the dorsal fin is variable ranging from low and stubby with a broad base to high and falcate (curved).
4. The shell is very long and narrow, falcate, fibrous, and distinctly exhibiting the small septa as they occur in the genus Caprina.
5. Many of the specimens are strongly falcate and appear to be pinnules rather than leaves; they often have longitudinal wrinkles that appear to result from compression of a thick lamina.
6. Macroconidia are falcate and have three or four septa.

```
falcate: Botany Zoology, 0.10162592 Anatomy  FAIL
falcate: Botany Zoology, 0.060674522 Anatomy  FAIL
falcate: Botany Zoology, 0.060319785 Anatomy  FAIL
falcate: Botany Zoology, 0.08161455 Anatomy  FAIL
falcate: Botany Zoology, 0.056197796 Entomology  FAIL
falcate: Botany Zoology, 0.08503878 Entomology  FAIL

0.0 % of domains are the same.
100.0 % of domains are different.
```

Teniendo en cuenta que *Entomology* puede tener relación con *Zoology*, podríamos decir que más o menos ha acertado en los dos últimos casos. Sin embargo, el dominio que más peso tiene es *Anatomy*. De nuevo, el etiquetado de Ask2Transformers tiene sentido por algunas palabras que contienen las frases. Al contrario de los casos que hemos visto hasta ahora, a pesar de no coincidir el dominio principal, *Anatomy* se acerca más al significado de *falcate* que *Geometry*, que ha sido la etiqueta propuesta a la definición en secciones anteriores.

***solifuge**: A sun spider.*

1. They're actually solpugids, or solifugids (aka camel spiders, aka wind scorpions).
2. Chomping down with oversize jaws, a wind scorpion (also known as a camel spider, sun spider, or solifugid) lunches on a lizard in California's Mojave Desert.
3. Despite the name and appearance, camel spiders are actually solifugids which, unlike spiders, do not have venom or silk glands.
4. Despite their fearsome appearance and their strong bite, solifugids are unlikely to harm humans.

```
solifuge: Zoology, 0.15893018 Nautical  FAIL
solifuge: Zoology, 0.22370353 Nautical  FAIL
solifuge: Zoology, 0.07304086 Nautical  FAIL
solifuge: Zoology, 0.040927287 Logic  FAIL

0.0 % of domains are the same.
100.0 % of domains are different.
```

En este caso, el dominio que más se repite, *Nautical*, no tiene ninguna relación con la palabra. En el apartado 5.3.2 en el que utilizábamos 90 dominios, el resultado era *Entomology*, por lo que se acercaba más al significado. Sin embargo, en la sección 5.3.3 con dominios agrupados, se le asignaba el dominio *Meteorology*, suponemos que por la palabra *sun* en la definicion.

***stannic**: Of tin with a valency of four; of tin(IV).*

1. This book presents the background science and technology of the stannic oxide gas sensor, along with practical information about its applications.
2. Tin oxide or stannic oxide is commonly used as an opacifier in ceramic glazes.

3. Single crystals of stannic oxide heavily doped with antimony have been grown from tin vapour and oxygen at 1450°c.

```
stannic: Chemistry, 0.04337446 Greek_Mythology  FAIL
stannic: Chemistry, 0.05980777 Anatomy  FAIL
stannic: Chemistry, 0.04661188 Law  FAIL

0.0 % of domains are the same.
100.0 % of domains are different.
```

Los dominios devueltos por Ask2Transformers con estos ejemplos como entrada no tienen nada que ver con el significado de *stannic*. En esta prueba no cabe duda de que el resultado de clasificar la definición es mejor, ya que con 90 dominios el resultado es *Mineralogy* y con 47 agrupados, *Chemistry*.

***tabes**: Emaciation.*

1. If no other symptoms of tabes can be found, it is an eye lesion.
2. The average duration of tabes can be placed at from 10 to 20 years.

```
tabes: Medicine, 0.048352383 Entomology  FAIL
tabes: Medicine, 0.060007412 Entomology  FAIL

0.0 % of domains are the same.
100.0 % of domains are different.
```

La definición de *tabes* es muy corta. Aun así, aunque los ejemplos sí que ayuden a entender mejor el significado de la palabra, la clasificación del programa tampoco es acertada. En las secciones anteriores la definición se enlazaba con los dominios *Accounting* y *Ecology*, por lo que tampoco era mejor el resultado.

5.4. Conclusiones

Después de haber realizado las diferentes pruebas, podemos sacar algunas conclusiones como:

1. La clasificación de dominios hecha por Oxford es muy amplia y concreta, lo cual complica que otro clasificador de dominios acierte y coincida su etiquetado con el del diccionario. En parte esto se debe a la alta cantidad de "subdominios"que hay (*American_Football* $\rightarrow$ *Sport*)

2. En base al punto anterior, cuanto menor sea el listado de dominios y cuanto menos parecidos sean los dominios entre sí (o menos "subdominios"haya), el clasificador tiene más posibilidades de acertar con el etiquetado.

3. Si hacemos un *dataset* con los dominios que más se repiten en Oxford, también ayudará al clasificador a acertar más en su etiquetado.

4. Clasificar las frases de ejemplo de una entrada en vez de su definición es muy arriesgado, ya que los ejemplos pueden contener palabras de contextos distintos al del significado que pueden desviar la atención del clasificador.

5. Hacer un clasificador de dominios es una tarea muy complicada aún, y aunque ya se puedan conseguir resultados óptimos en muchos casos, todavía se puede mejorar más.

6. CAPÍTULO

Conclusiones y trabajo futuro

6.1. Conclusiones

Los objetivos principales de este proyecto eran adquirir el conocimiento léxico del diccionario en línea de Oxford y aplicar y evaluar un clasificador de dominios zero-shot con modelos de lenguaje pre-entrenados. Gracias al trabajo realizado, hemos podido alcanzar ambos objetivos. En concreto:

- Hemos generado código dedicado a la estructura concreta de la página web del diccionario de Oxford para obtener primero todas las palabras que contiene, y después todas las palabras junto con sus a veces múltiples definiciones, ejemplos, dominios y sinónimos entre otros atributos.
- Hemos analizado y clasificado los dominios del diccionario. Después, hemos comprobado el funcionamiento del clasificador de dominios zero-shot Ask2Transformers con definiciones y ejemplos de Oxford. Hemos realizado distintas pruebas con distintos *dataset* de dominios, definiciones y frases de ejemplo y hemos analizado cómo cambian los resultados. Tras finalizar las pruebas, hemos llegado a la conclusión de que la clasificación de dominios hecha por Oxford es muy amplia y concreta, lo cual complica que otro clasificador de dominios acierte y coincida su etiquetado con el del diccionario. De todos modos, el clasificador Ask2Transformers funciona mejor cuanto menor sea el listado de dominios, más diferentes sean los dominios entre sí y se usen los más frecuentes. En estos casos, la tasa de acierto es superior

al 50 % y a veces superior al 75 %. Es mejor utilizar una palabra con su definición como entrada que los ejemplos directamente, ya que el contexto de las frases de ejemplo puede provocar malinterpretaciones en el clasificador. En general, a pesar de los múltiples avances de los últimos años en el área del PLN, aún queda mucho por seguir mejorando hasta llegar a la perfección absoluta.

Por último, cabe destacar que al inicio de este proyecto mi conocimiento sobre el PLN se reducía a lo visto en la asignatura de Procesamiento del Lenguaje o *Hizkuntzaren Prozesamendua* (HP). El desarrollo de este TFG ha despertado mi interés en seguir aprendiendo sobre esta disciplina tan desconocida para algunos y tan necesaria para todos.

6.2. Trabajo futuro

Si bien en los últimos años se han realizado avances espectaculares, los fundamentos teóricos del Procesamiento del Lenguaje Natural se encuentran todavía en estado de desarrollo. Existen infinitas posibilidades para contribuir en estos avances, y lo mismo pasa con la adquisición de conocimiento léxico en este área. A continuación mostramos cómo se podría hacer una contribución en relación a este proyecto.

1. Existen muchas otras formas de probar el programa **Ask2Transformers** con dominios del diccionario de Oxford y así seguir comparando sus resultados para hacer mejoras en él. Una de ellas puede ser hacer uso de los sinónimos en lugar de las definiciones para evaluar el etiquetado del programa. Algo parecido a lo que ya hemos hecho con las frases de ejemplo.

2. Al igual que hemos hecho con Ask2Transformers, también se pueden probar diferentes modelos de lenguaje. En la plataforma de **Hugging Face**[1] existen varios modelos monolíngües, como por ejemplo:

   - RoBERTa: A Robustly Optimized BERT Pretraining Approach [2] [Liu et al., 2019]. Dadas dos frases, las clasifica como *Contradiction*, *Neutral* o *Entailment* dándoles un peso a cada una de las tres etiquetas.
   - NLI-based Zero Shot Text Classification [3]. Dado un texto y una lista de dominios, asigna un peso a cada dominio dándole el mayor peso al que tenga más

[1] https://huggingface.co/
[2] https://huggingface.co/roberta-large-mnli
[3] https://huggingface.co/facebook/bart-large-mnli

relación con el texto, y el menor peso al dominio que menos. Es parecido a Ask2Transformers.

- DeBERTa: Decoding-enhanced BERT with Disentangled Attention [4] [5] [6] [7] [8] [He et al., 2021]. Dadas dos frases, las clasifica como *Contradiction*, *Neutral* o *Entailment* dándoles un peso a cada una de las tres etiquetas. Mejora los modelos BERT y RoBERTa superándolos en la mayoría de las tareas de NLU con datos de entrenamiento de 80 GB.

Y alguno multilingüe:

- Xlm-roberta-large-xnli [9]. Este modelo toma xlm-roberta-large y lo ajusta en una combinación de datos NLI en 15 idiomas. Está diseñado para ser utilizado para la clasificación de texto de zero-shot. Al igual que Ask2Transformers, se dedica a asignar dominios con sus pesos a una frase, solo que la frase puede estar en un idioma distinto al de los dominios.

3. **WordNet**[10] es una gran base de datos léxica del inglés. Todos los sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos cognitivos llamados *synsets*, cada uno expresando un concepto distinto. Se asemeja superficialmente a un tesauro, ya que agrupa las palabras en función de sus significados. Por tanto, podría enlazarse el diccionario de Oxford con WordNet, ya que su estructura lo convierte en una herramienta útil para la ligüística computacional y el procesamiento del lenguaje natural.

4. El resultado de este proyecto deja a nuestra disposición el contenido de un diccionario en línea. Esta información podemos traducirla haciendo uso de traductores automáticos y así generar diccionarios en otros idiomas.

5. Al igual que hemos adquirido el conocimiento léxico del diccionario de Oxford, podemos hacer lo mismo con otros diccionarios o tesauros en línea, como por ejemplo:

   - https://www.merriam-webster.com/browse/thesaurus/
   - https://www.merriam-webster.com/browse/dictionary/

[4] https://huggingface.co/microsoft/deberta-base-mnli
[5] https://huggingface.co/microsoft/deberta-large-mnli
[6] https://huggingface.co/microsoft/deberta-xlarge-mnli
[7] https://huggingface.co/microsoft/deberta-v2-xlarge-mnli
[8] https://huggingface.co/microsoft/deberta-v2-xxlarge-mnli
[9] https://huggingface.co/joeddav/xlm-roberta-large-xnli
[10] http://wordnetweb.princeton.edu/perl/webwn

- https://www.dictionary.com/list/a

El siguiente enlace muestra un listado de 129 diccionarios en línea con los que se podría repetir el proceso: https://onelook.com/?d=all_gen

Anexos

Bibliografía

[A. and R, 1972] A., C. and R, Q. M. (1972). Experiments on semantic memory and language comprehension. *Cognition in Learning and Memory. Wiley, New York.*

[Castillo and Rigau, 2013] Castillo, M. and Rigau, G. (2013). Adquisición automática de conocimiento léxico con wordnet. *Universidad del País Vasco grado de PhD*.

[Chang et al., 2018] Chang, T.-Y., Chi, T.-C., Tsai, S.-C., and Chen, Y.-N. (2018). xsense: Learning sense-separated sparse representations and textual definitions for explainable word sense networks. *arXiv preprint arXiv:1809.03348*.

[Gadetsky and Vetrov, 2018] Gadetsky, A.; Yakubovskiy, I. and Vetrov, D. (2018). Conditional generators of words definitions. in proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers), 266–271. *arXiv preprint arXiv:1806.10090*.

[He et al., 2021] He, P., Liu, X., Gao, J., and Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[Sainz and Rigau, 2020] Sainz, O. and Rigau, G. (2020). Ask2transformers - zero shot domain labelling with pretrained transformers.

[SEPLN, 2020] SEPLN (2020). Procesamiento del lenguaje natural. *SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural)*.

[Wikipedia, 2020] Wikipedia (2020). Word embedding — wikipedia, la enciclopedia libre. [Internet; descargado 5-septiembre-2021].

[Wikipedia, 2021a] Wikipedia (2021a). Bert (modelo de lenguaje) — wikipedia, la enciclopedia libre. [Internet; descargado 5-septiembre-2021].

[Wikipedia, 2021b] Wikipedia (2021b). Procesamiento de lenguajes naturales — wikipedia, la enciclopedia libre. [Internet; descargado 5-septiembre-2021].

[Wikipedia, 2021c] Wikipedia (2021c). Traducción automática — wikipedia, la enciclopedia libre. [Internet; descargado 5-septiembre-2021].

[Wikipedia contributors, 2021a] Wikipedia contributors (2021a). Longman dictionary of contemporary english — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Longman_Dictionary_of_Contemporary_English&oldid=1041081619. [Online; accessed 5-September-2021].

[Wikipedia contributors, 2021b] Wikipedia contributors (2021b). Machine-readable dictionary — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Machine-readable_dictionary&oldid=1030161087. [Online; accessed 5-September-2021].

[Wikipedia contributors, 2021c] Wikipedia contributors (2021c). Wordnet — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=WordNet&oldid=1035371095. [Online; accessed 5-September-2021].