

Listas de contenidos disponibles en [CienciaDirecta](#)

Sistemas Expertos con Aplicaciones

revista Página de inicio: www.elsevier.com/locate/eswa

Revisar

Resumen abstractivo: una visión general del estado del arte

Som Gupta ^{a,*}, S. K. Gupta ^b^aAKTU Lucknow, UP 226031, India^bBIET Jhansi, UP 284128, India

información del artículo

Historial del artículo:

Recibido el 3 de septiembre de 2018 Revisado el
28 de octubre de 2018 Aceptado el 6 de
diciembre de 2018 Disponible en línea el 7 de
diciembre de 2018

Palabras clave:

Resumen abstracto
Búsqueda de conceptos
Resumen basado en semántica
Resumen basado en ontología
Aprendizaje profundo

resumen

El resumen, es reducir el tamaño del documento conservando el significado, es una de las áreas más investigadas entre la comunidad de procesamiento de lenguaje natural (NLP). Las técnicas de resumen, en función de si se consideran las oraciones exactas tal como aparecen en el texto original o si se generan nuevas oraciones utilizando técnicas de procesamiento del lenguaje natural, se clasifican en técnicas extractivas y abstractas. El resumen extractivo ha sido un tema muy investigado y ha llegado a su etapa de madurez. Ahora la investigación se ha desplazado hacia la síntesis abstractiva. Las complejidades subyacentes al texto en lenguaje natural hacen que el resumen abstracto sea una tarea difícil y una tarea desafiante.

Este artículo presenta una revisión exhaustiva de los diversos trabajos realizados en resumen abstracto. campo de rización. Para ello, hemos seleccionado los artículos recientes sobre este tema de Elsevier, ACM, IEEE, Springer, ACL Anthology, Cornell University Library y Google Scholar. Los artículos se clasifican según el tipo de técnica abstractiva utilizada. El documento enumera los diversos desafíos y analiza la dirección futura de la investigación en este campo. Junto con estos, hemos identificado las ventajas y desventajas de varios métodos utilizados para el resumen abstracto. También hemos enumerado las diversas herramientas que han utilizado o desarrollado los investigadores para el resumen abstracto. El documento también analiza las técnicas de evaluación que se utilizan para evaluar los resúmenes abstractivos.

© 2018 Elsevier Ltd. Todos los derechos reservados.

1. Introducción

El crecimiento exponencial de la información debido a la popularidad del entorno web ha llevado a la necesidad de resúmenes automáticos para reducir el esfuerzo y el tiempo mientras se encuentra la información concisa y relevante de acuerdo con la consulta. El resumen consiste en reducir el contenido del texto conservando el significado del texto. Varias formas de escribir lo mismo han hecho de este tema un tema interesante entre los investigadores. Se ha trabajado mucho en el área de resumen automático en los últimos años. Según cómo se seleccione y organice el contenido en el resumen, las técnicas de resumen se clasifican en técnicas extractivas y abstractas. El resumen extractivo consiste en encontrar las oraciones más destacadas del texto considerando las características estadísticas y luego organizando las oraciones extraídas para crear el resumen. El resumen abstractivo, por otro lado, es una técnica en la que el resumen es generado por

generar oraciones nuevas ya sea reformulando o usando las palabras nuevas, en lugar de simplemente extraer las oraciones importantes. La representación interna del texto se crea mediante el análisis de la información semántica sobre el texto y, con un análisis y razonamiento profundos, se generan nuevas oraciones a partir del texto original (Rachabatuni, 2017). Según el número de documentos que se consideran para el resumen, se clasifican en resumen de un solo documento y de varios documentos. En función de cuanta información se va a resumir, se clasifican en genéricos, donde se obtiene el resumen de todo el texto y resumen centrado en la consulta, donde solo se obtiene el resumen de acuerdo al contexto especificado por el usuario. Y, en función del contenido, se clasifican en indicativos, donde el foco está en contar de qué trata el texto y resumen informativo, donde el contenido principal del texto se extrae mediante el análisis del texto original. Ha habido mucha investigación en resúmenes extractivos, pero los resúmenes extractivos carecen de cohesión, legibilidad y otros factores de calidad debido a la presencia de problemas de anáforas colgantes. También, Yao, Wan y Xiao, (2017).

*Autor correspondiente.

Correos electrónicos: somi.11ce@gmail.com (S. Gupta), guptask_biet@rediffmail.com (SK Gupta).

URL: <http://www.aktu.ac.in> (S. Gupta)

De nuestra encuesta, hemos encontrado que la mayoría de los trabajos son sobre resumen extractivo debido a la simplicidad para implementarlo. Mientras que la complejidad del procesamiento del lenguaje natural hace que el resumen abstracto sea una tarea desafiante. Pero ahora la investigación sobre resúmenes extractivos se ha estancado ya que han alcanzado el máximo rendimiento (Mehta, 2016) y, por lo tanto, la atención se ha desplazado más hacia el resumen abstracto y la fusión de técnicas extractivas y abstractas. La necesidad de tiempo es producir resúmenes de alta calidad que sean gramaticalmente correctos, legibles, coherentes, concisos y ricos en información. El resumen abstracto ayuda a resolver el problema de la anáfora colgante y, por lo tanto, ayuda a generar resúmenes legibles, concisos y cohesivos. El resumen abstracto ayuda a reducir el tamaño de la oración, ya que utiliza la fusión para fusionar las oraciones, por lo que ayuda a lograr una mayor ausencia de redundancia en el resumen en comparación con los resúmenes extractivos, donde incluso la parte no relevante de la oración también se incluye debido al hecho de que extrae las oraciones y las ordena. Los resúmenes abstractivos también dan mayor precisión que los resúmenes extractivos.

Hemos utilizado la búsqueda basada en citas para identificar los artículos relevantes. Para garantizar la buena calidad de los documentos para nuestro propósito de revisión, hemos utilizado IEEE,¹ACM,²Salador,³antología de ligamento cruzado anterior,⁴ biblioteca de la universidad de cornell,⁵Elsevier⁶y Google Académico.⁷

La mayoría de los trabajos en el campo del resumen abstracto se centran en componentes como el análisis sintáctico, la resolución de correferencias, la construcción y fusión de gráficos semánticos, la generación de lenguaje natural, las cadenas léxicas y la semántica distribucional (Yao et al., 2017). La mayoría de las técnicas abstractivas involucran compresión de oraciones, fusión (Belkebir y Guessoum, 2016; Nayeem y Chali, 2017) o revisión para generar el resumen final a partir de las oraciones seleccionadas.

Según se considere la estructura del texto o la semántica, las técnicas de resumen abstracto se clasifican en técnicas basadas en la estructura y técnicas basadas en la semántica. Aparte de la estructura y la semántica, hoy en día el aprendizaje profundo ha surgido como una nueva técnica para modelar el problema de resumen abstracto que puede capturar tanto la información estructural como la semántica del texto.

A partir de la revisión de la literatura, hemos encontrado que la cantidad de artículos que describen las técnicas de resumen extractivo en detalle es enorme, pero hay muy pocos artículos y en su mayoría están desactualizados en el campo del resumen abstracto que enumera los diversos trabajos recientes realizados en este campo. de manera categorizada y detallada. Hay documentos separados disponibles para cada tema individual, como compresión de oraciones, fusión de oraciones, resumen mediante aprendizaje profundo, pero no hay ningún documento que los enumere juntos. Entonces, para abordar este problema, hemos enumerado los diversos trabajos realizados hasta mediados de 2018 y los clasificamos en la jerarquía de clasificación predefinida. Nuestro estudio permitirá sistematizar y potenciar el conocimiento en este campo de investigación. También hemos discutido los componentes populares del sistema de resumen abstracto.

En este artículo, discutimos las tendencias de investigación en el campo de la síntesis abstractiva. Damos una visión general de varias técnicas de resumen abstracto disponibles. Discutimos las herramientas que se han utilizado para crear resúmenes abstractos. También discutimos las medidas de evaluación utilizadas para evaluar estos resúmenes. Tenemos

identificó los desafíos que se encuentran en el campo de la síntesis abstractiva y discutió las tendencias futuras en esta área de investigación.

Nuestro artículo responde a las siguientes preguntas:

- Cómo se clasifican las técnicas de resumen abstracto
- Cuáles son los diversos trabajos recientes realizados en este campo según la técnica de resumen
- ¿Cuáles son las diversas herramientas que se han utilizado para crear los resúmenes abstractivos?
- Cómo varían los resultados entre las diversas técnicas
- ¿Cuáles son los famosos conjuntos de datos que se han utilizado para evaluar y realizar la tarea de resumen abstracto?
- ¿Cuáles son los diversos desafíos y problemas de investigación abiertos que se encuentran en este campo de investigación?

El documento está organizado de la siguiente manera: -Sección 2 describe las diversas técnicas de resumen abstracto. Sección 3 analiza algunos de los pasos comunes que están involucrados al crear los resúmenes. Sección 4 analiza las diversas herramientas utilizadas para realizar resúmenes abstractivos. Sección 5 señala los desafíos y la dirección futura para los nuevos investigadores. Y finalmente la conclusión.

2. Enfoques existentes para el resumen abstracto

Moratanch y Chitrakala (2016), Kan (2014) y Dineshnath y S. Saraswathi (2018) han clasificado los enfoques de resumen abstracto ampliamente en basados en la estructura y basados en la semántica. Después de examinar los trabajos de investigación de resumen abstracto, hemos agregado un enfoque más, el aprendizaje profundo con redes neuronales. Los enfoques basados en la estructura son aquellos en los que la información importante del texto se completa en la estructura predefinida para crear resúmenes abstractos. Los enfoques basados en estructuras se dividen en métodos basados en árboles, en plantillas, en ontologías, en frases de guía y cuerpo, en gráficos y en reglas, según la estructura utilizada para crear resúmenes. Mientras que los enfoques basados en la semántica son aquellos que toman el documento de texto como entrada, crean la representación semántica del texto y luego alimentan esta representación al sistema de generación de lenguaje natural para crear el resumen abstracto final. Figura 1 es la clasificación de los enfoques de resumen abstracto. En las siguientes subsecciones, hemos discutido los diversos métodos de resumen abstracto.

2.1. Enfoque basado en la estructura

Los enfoques basados en estructuras encuentran la información más importante del texto y luego usan plantillas, reglas, árboles, ontología, etc. para crear resúmenes abstractos. Se utilizan principalmente junto con enfoques extractivos, basados en la semántica o en el aprendizaje profundo. Algunos de los ejemplos son como Ganesan, Zhai y Han (2010) han utilizado plantillas junto con el enfoque basado en gráficos, Bartakke, Sawarkar y Gulati (2016) reglas usadas y ontología junto con los grafos semánticos, Li (2015) plantillas usadas junto con el enfoque basado en la semántica, Wang y Ling (2016) utilizó redes neuronales junto con las plantillas para crear los resúmenes abstractos. A veces, estos enfoques también se utilizan como el primer paso para el preprocesamiento de texto como Nguyen y Phan (2009) extrajo las frases clave importantes del texto usando ontología que luego se puede combinar con algunos otros enfoques para crear el resumen abstracto. A continuación se muestra la lista de varios métodos que se utilizan para crear resúmenes abstractos teniendo en cuenta la estructura del texto.

¹<https://ieeexplore.ieee.org/>.

²<https://dl.acm.org/>.

³<https://link.springer.com>.

⁴<http://aclweb.org/antologia/>.

⁵<https://arxiv.org/>.

⁶<https://www.elsevier.com>.

⁷<https://scholar.google.co.in/>.

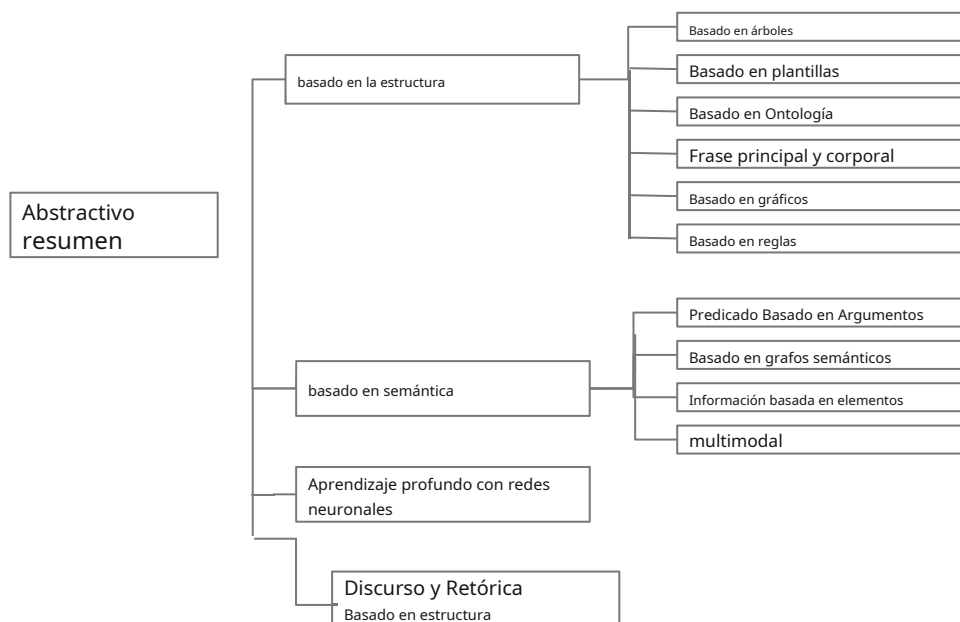


Figura 1. Clasificación de la técnica de resumen abstracto.

2.1.1. Métodos basados en árboles

En este enfoque, primero se extrae el texto importante a considerar para el resumen. Luego, las oraciones similares se identifican a partir de este texto mediante el uso de un analizador superficial. Las oraciones similares de este texto se rellenan luego en la estructura en forma de árbol. El procesamiento de los árboles se realiza mediante la linealización (conversión de árboles en cadenas), la búsqueda de una estructura predicado-argumento o la fusión de oraciones para crear los resúmenes abstractos finales (Sunitha, Jaya y Ganesh, 2016). Hay varios algoritmos que nos ayudan a realizarlos. Por ejemplo, el algoritmo de intersección de tema de contenido es uno que se usa para determinar las frases comunes usando una estructura de argumento de predicado, encontrando el árbol base (Barzilay y McKeown, 2005) al encontrar el centroide del árbol de dependencia y luego aumentar estos árboles base para obtener los subárboles también es otra forma de evaluar la oración. Se obtienen frases comunes y luego se les agrega alguna información o se alimentan a los sistemas de generación de lenguaje natural para crear las nuevas oraciones y luego se organizan juntas para crear un resumen.

El árbol de dependencia es la estructura de datos más popular, utilizada para representar el texto en forma de árbol. Barzilay y McKeown (2005) utilizó la generación de texto a texto para crear los resúmenes informativos. Representaron las oraciones usando árboles de dependencia y encontraron la información común entre las oraciones procesando los árboles. Calcularon la red de fusión al encontrar la intersección de los subárboles y luego usaron recorridos de árboles en ellos para producir la oración final. Una de las limitaciones de este enfoque es que no puede capturar las conexiones entre las oraciones sin encontrar la frase interseccionada entre las oraciones. Yousfi-Monod y Prince (2008) desarrolló un enfoque llamado CoLIN, basado en la linealización y la poda del árbol de dependencias, manteniendo la información semántica, el contenido de la información, la corrección gramatical y la coherencia en el resumen. Realizaron el análisis lingüístico profundo del texto y se centraron en la heurística lingüística.

Kurisinkel, Zhang y Varma (2017) usó los árboles de dependencia parcial que se construyen a partir de árboles de dependencia sintáctica analizando el texto junto con la linealización sintáctica basada en la recombinación y la transición para crear resúmenes abstractos de varios documentos.

El uso de sistemas de generación de lenguaje natural para la creación de resúmenes ayuda a lograr la fluidez y la corrección gramatical. Pero en su mayoría no consideran el contexto al encontrar las frases importantes que se incluirán en el resumen (Kasture1, Yargal, Singh, Kulkarni y Mathur, 2014). Se basan principalmente en el análisis y la alineación de los árboles de análisis (Khan et al., 2018). Los árboles de dependencia se construyen con la ayuda de analizadores. Por lo tanto, el rendimiento de estos métodos depende en gran medida de los analizadores disponibles, lo que limita su eficiencia. Se centran más en la sintaxis que en la semántica.

2.1.2. Métodos basados en plantillas

En los métodos basados en plantillas, los fragmentos se extraen mediante palabras clave o pistas; y los fragmentos extraídos se rellenan en plantillas para formar el resumen final. Debido a que la estructura está predefinida, ayuda a crear resúmenes concisos y coherentes. Se basan en un profundo análisis sintáctico y semántico del texto. Es uno de los métodos más populares para crear resúmenes de varios documentos. Este método funciona bien cuando el texto está estructurado de alguna manera. Pero, debido a que las reglas y los patrones se definen manualmente, este método consume mucho tiempo y también requiere mucho esfuerzo manual (Alshaina, John y Nath, 2017).

Harabagiu et al. (2001) usó el resumen basado en plantillas y creó un sistema llamado GISTEXTER para crear los resúmenes multidocumento donde las plantillas se completan siguiendo los patrones y las reglas. Carenini, Ng y Pauls (2012) utilizó las plantillas para generar los resúmenes en lenguaje natural para el corpus evaluativo en su sistema SEA. Embar, Deshpande, K, Jain y Kallimani (2013) usó el método de plantilla basado en dominio para crear el resumen abstracto basado en Kannada. Utilizaron las reglas de extracción de información hechas a mano para extraer información para llenar las plantillas. Oya, Mehdad, Carenini y Ng (2014) propuso una técnica de fusión de oraciones múltiples para crear plantillas abstractas. Para crear plantillas, han utilizado las frases nominales junto con el etiquetado POS y los hiperónimos. Después de encontrar plantillas, las han agrupado extrayendo verbos raíz. Usaron gráficos de palabras para la fusión de plantillas para generar el resumen final. Gerani, Mehdad, Carenini, Ng y Nejat (2014) también han utilizado plantillas para generar los resúmenes utilizando las frases seleccionadas obtenidas al considerar la estructura discursiva del texto.

2.1.3. Métodos basados en ontología

Una ontología se puede considerar como la colección de entidades y la relación entre ellas. Una ontología junto con el conjunto de instancias individuales de clase constituye la base de conocimiento (Mohan, Sunitha, Ganesha y Jaya Dr., 2016). Las clases son la parte más importante de una ontología y representan los conceptos. Una ontología define el conjunto de vocabulario y también contiene un acuerdo de términos, lo que ayuda a eliminar cualquier tipo de ambigüedad al encontrar los conceptos. Las ontologías generalmente son creadas por expertos en el dominio y se utilizan principalmente para extraer los conceptos y las relaciones del texto. Estos conceptos se utilizan luego para crear los resúmenes (Mohan et al., 2016). Uno de los ejemplos de una ontología específica de dominio es "WordNet". Las ontologías ayudan a compartir un conocimiento común entre la comunidad de interés. Permiten la reutilización del conocimiento. Ayudan a separar el dominio y el conocimiento operativo y, por lo tanto, ayudan a realizar modificaciones fácilmente cuando cambia algo relacionado con el dominio. Se han utilizado en varios campos como el aprendizaje electrónico, el análisis de contenido relacionado con el usuario y el análisis de imágenes (Baralis, Cagliero, Jabeena, Fiori y Shah, 2013). Los sistemas abstractivos basados en ontologías extraen la información de las ontologías para crear resúmenes específicos para las necesidades del usuario.

Muchas veces el vocabulario en línea es limitado, la ontología ayuda en la representación adecuada del documento (Rananavare y Reddy, 2017). Incluso la representación semántica del contenido de la información se puede mejorar mucho con la ayuda de una ontología. La normalización del vocabulario se realiza para resolver la ambigüedad cuando hay más de un sinónimo de un concepto (Mohan et al., 2016).

La ontología se representa principalmente en una estructura jerárquica como el gráfico acíclico dirigido, etc. (Lloret, Teresa, Rom-Ferri y Palomara, 2013). Resource Development Framework (RDF) y Web Ontology Language (OWL) son dos lenguajes principales que se utilizan para crear una ontología. DBPedia (Bizer et al., 2009) es uno de los proyectos que ha estructurado los datos de wikipedia en forma de ontología. Las ontologías ayudan en la expansión de consultas y la búsqueda de consultas (Baralis et al., 2013). Tran, Cimiano, Rudolph y Studer (2007) usó las ontologías para construir su sistema de preguntas y respuestas donde usaron las ontologías para convertir la pregunta del usuario en la consulta del sistema, para que el sistema pueda comprender la consulta especificada por el usuario. Las ontologías ayudan a lograr la integridad del tema y la no redundancia en el resumen (Hípola, Senso, Leiva-Mederos, & Domínguez-Velasco, 2014).

Zhang, Cheng y Qu (2011) creó un enfoque para crear resúmenes personalizables basados en ontologías donde se puede especificar la longitud del resumen y las preferencias de navegación. Se construyeron gráficos RDF y se obtuvieron las puntuaciones de prominencia de las oraciones RDF para encontrar las oraciones importantes desde el punto de vista del usuario. Hennig, Umbrath y Wetzker (2008a) crearon un sistema de resumen de texto basado en ontología donde aplicaron el clasificador jerárquico que mapeó las oraciones en los nodos de una ontología predefinida. Lee, Chen y Jian (2003) creó un enfoque basado en ontologías para el resumen de noticias en chino mediante el uso de la ontología predefinida creada para artículos de noticias. Usaron la ontología para inferir los eventos de las noticias. Hipola et al. (2014) utilizaron la ontología en su sistema llamada Texminer para extraer las oraciones importantes con estructura retórica, que resume la información de puertos y costas. La ontología ayudó a contener todos los conceptos importantes de los temas de ingeniería costera. La ontología ayudó a capturar la información semántica de los conceptos para facilitar el resumen en su sistema. Dolfo Lozano-Tello y Gómez-Pérez (2004) creó una métrica llamada OntoMetric para elegir la ontología adecuada de acuerdo con las especificaciones del requisito. Nguyen y Phan (2009) usó la ontología de Wikipedia para extraer las frases clave que son la representación lingüística del texto, del texto vietnamita. Estas frases clave se pueden utilizar para resumir documentos. Baralis et al. (2013) presentó un resumen abstracto multidocumento que utilizó la ontología de Yago para

selección de oraciones por la identificación de entidades, conceptos y desambiguación.

2.1.4. Métodos de frases principales y corporales

Aquí se utiliza la técnica de inserción y sustitución de información importante con frases ricas en la oración principal del cuerpo del texto, que se denominan factores desencadenantes (Sunitha et al., 2016). La creación de oraciones gramaticalmente correctas mediante el uso de este método sigue siendo un problema. Tanaka, Kinoshita, Kobayakawa, Kumano y Kato (2009) creó los resúmenes de frases principales y del cuerpo del resumen de transmisión. Revisaron los resúmenes modificando la oración principal mediante la inserción y sustitución de frases. Modificaron la oración principal agregando información como "quién", "dónde", "cuándo", "cómo", etc.

2.1.5. Métodos basados en gráficos

Word Graph es la estructura de datos de gráficos más popular que se utiliza para representar el texto en estos métodos. Las oraciones se fusionan para crear los resúmenes abstractos. La fusión de oraciones ayuda a eliminar la parte redundante de la oración y así lograr la no redundancia. Los gráficos de palabras funcionan asumiendo que habrá muchas oraciones similares en el texto y esta similitud ayudará a fusionar las oraciones. Pero no es necesario que sea fácil encontrar frases semejantes.

Barzilay y Lee (2003a) utilizó la paráfrasis para generar los resúmenes en lenguaje natural mediante la fusión de oraciones. Usaron la estructura basada en gráficos para encontrar oraciones estructuralmente similares, luego identificaron el texto que se parafrasea entre sí y luego generaron la oración.

Katja (2010) creó los resúmenes gramaticalmente correctos e informativos encontrando el camino más corto en los gráficos de palabras. Pero como clasificaron sus oraciones sobre la base de la información, carecían de la calidad lingüística. Mehdad, Carenini, Tompa y Ng (2013) amplió la obra de Katja (2010) creando los gráficos de vinculación para eliminar la información redundante junto con los gráficos de palabras para encontrar las oraciones relevantes del texto informal de las transcripciones de las reuniones, para crear los resúmenes abstractos. Usaron WordNet para encontrar las relaciones entre las palabras y usaron esta información para fusionar los nodos del gráfico. El contenido de la información y la fluidez gramatical se eligieron como los factores para decidir los mejores caminos del gráfico y luego se utilizó la generalización y la agregación para generar el resumen abstracto.

Ganesan et al. (2010) creó resúmenes abstractos informativos, legibles, concisos y bien formados para opiniones redundantes mediante el uso de una estructura de datos basada en gráficos llamada Word Graph. Su enfoque no requiere conocimiento del dominio y utiliza PNL superficial. Su gráfico de opinión captura la redundancia y ayuda a descubrir las nuevas oraciones al identificar los enlaces léxicos.

Le y Le (2013) creó los resúmenes extractivos y luego aplicó gráficos de palabras para reducir la oración y combinar las oraciones. Niu, Chen, Zhao, Sun y Atiquzzaman (2017) usó los gráficos de fragmentos que se basan en el gráfico de palabras para reducir el tamaño del gráfico. Oyá et al. (2014) usó el gráfico de palabras para fusionar las plantillas generadas para las conversaciones de las reuniones.

Banerjee, Mitra y Sugiyama (2017) usó los gráficos de palabras junto con la programación lineal entera (ILP) para crear los resúmenes abstractos de varios documentos. Primero identificaron los documentos más importantes entre los documentos que se resumirían con la ayuda de LexRank, el puntaje de similitud del coseno y el puntaje general de similitud de la colección de documentos. Luego, crearon grupos de oraciones similares entre los documentos importantes. Las rutas más cortas se obtienen mediante la creación de gráficos de palabras y el modelo ILP se aplica para encontrar las oraciones con la máxima información y legibilidad. ILP ayuda a minimizar la redundancia en el resumen.

2.1.6. Métodos basados en reglas

En este método, las reglas y categorías se introducen en el sistema para encontrar los candidatos significativos que luego se utilizan para crear el resumen. Aquí, el documento de texto se clasifica primero de acuerdo con los términos y conceptos presentes en ellos. Luego se formulan las preguntas según el dominio del documento de texto y luego se extraen las respuestas del documento. Las preguntas pueden ser como: dónde ocurrió el evento, cuándo ocurrió el evento, quién hizo el evento, cuál fue el impacto del evento, etc. Estas preguntas generalmente se responden encontrando la parte del discurso de los términos y conceptos en el texto. Al responder estas preguntas, se introducen en un patrón que luego ayuda a crear el resumen abstracto final (Kasture¹ et al., 2014). Pero nuevamente, como en el enfoque basado en plantillas, aquí las reglas se escriben manualmente y, por lo tanto, conducen a una pérdida de tiempo.

Tabla 2 enumera las ventajas y desventajas de los enfoques basados en estructuras en la página 15.

2.2. Enfoque basado en la semántica

Aquí primero se obtiene la representación semántica del texto encontrando los elementos de información, la estructura predicado-argumento o creando gráficos semánticos. Luego, esta representación se alimenta al sistema de generación de lenguaje natural y mediante el uso de las frases nominales y verbales (Dineshnath y S. Saraswathi, 2017), se crea el resumen abstracto final. El algoritmo de intersección de temas de contenido también se utiliza con los gráficos para determinar las frases comunes mediante el uso de una estructura de argumento-predicado.

2.2.1. Métodos basados en elementos de información (INIT)

Genest y Lapalme (2011) usó el elemento de información, la unidad más pequeña de información coherente en el texto para encontrar los resúmenes abstractos. Con elemento de información, se entienden las entidades, sus características y las relaciones o propiedades entre ellas. Se basaron en el modelado de roles semánticos, la desambiguación, el análisis de co-referencia, el análisis de similitud y el análisis de lógica de predicados para encontrar los elementos de información. Usaron tríos de sujeto, verbo y objeto junto con la información de tiempo y ubicación para crear las oraciones de resumen. El marco general para los métodos basados en INIT contiene cuatro módulos, a saber, recuperación de elementos de información donde se extraen trillizos mediante un analizador, generación de oraciones mediante el uso de un generador de lenguaje, selección de oraciones mediante la búsqueda de las oraciones mejor valoradas, que se calcula sobre la base de factores como la frecuencia del documento, Kasture¹ et al., 2014).

2.2.2. Enfoque basado en predicado-argumento

La estructura predicado-argumento significa los verbos, el sujeto y el objeto, etc. de la oración. Se obtiene la estructura predicado-argumento de las oraciones para representarlas semánticamente. Luego, las estructuras semánticamente similares se encuentran a partir de ellas mediante el uso de medidas de similitud como la medida de distancia de edición, etc. Las estructuras semánticamente similares se fusionan mediante el uso de algunos métodos como K-means y agrupamiento jerárquico (Alshaina et al., 2017). Las características se extraen de las estructuras de predicado-argumento (PAS) y luego se puntúan. Para maximizar las puntuaciones de prominencia de las oraciones, se utilizan enfoques de optimización como la programación lineal entera (ILP). Los PAS de alta puntuación se seleccionan y se envían al sistema de generación de lenguaje para crear el resumen final. Pero generar las nuevas oraciones a partir de ellas es una tarea muy difícil (Li, 2015) de este método. Zhang, Zhou y Zong (2016) propuso un enfoque de resumen abstracto multilingüe mediante el uso de una estructura de predicado-argumento y fusionándolos mediante la programación lineal entera. Anotaron las estructuras de predicados y argumentos con modelos de roles semánticos para obtener los resúmenes. Para resolver la dependencia de la ontología escrita por humanos,

Alshaina et al. (2017) usó la estructura predicado-argumento junto con el modelo de rol semántico para encontrar la similitud semántica entre los argumentos y creó los resúmenes abstractos de varios documentos.

2.2.3. Métodos basados en grafos semánticos

Es una de las formas más populares de representar el texto sobre la base de las relaciones semánticas entre las oraciones del texto. Las propiedades semánticas incluyen las relaciones ontológicas y las relaciones sintácticas entre las palabras. Las relaciones ontológicas utilizan la propiedad de sinonimia, hiponimia, hiperonimia, etc., mientras que las relaciones sintácticas utilizan la propiedad de relación entre las palabras sobre la base de sujeto-objeto-verbo, es decir, representadas en términos de árbol de dependencia y árbol sintáctico (Joshi, Wang y McClean, 2018).

Aquí el documento se representa mediante un gráfico semántico. Los sustantivos y los verbos se representan como nodos del gráfico, mientras que las relaciones entre los nodos se representan por el borde. Khan et al. (2018) propuso el enfoque basado en grafos semánticos para la generación de resúmenes abstractos de múltiples documentos. Han, Lv, Hu, Wang y Wang (2016) propuso un modelo semántico basado en gráficos usando FrameNet donde cada oración se considera como un vértice y las relaciones semánticas entre las oraciones se representan como bordes. Framenet se usa para clasificar las oraciones y encontrar la relevancia entre las oraciones y la precisión se logra usando Wordnet. Los siguientes son algunos gráficos semánticos famosos que se han utilizado para crear resúmenes abstractos.

- Rich Semantic Graphs: Es una de las formas de representar la información semántica del texto en forma de gráficos. Los sustantivos y verbos del texto representan los nodos y la relación semántica y topológica entre ellos corresponden a los bordes. Los sustantivos y verbos del texto se obtienen con la ayuda de análisis sintáctico u ontología. Moawad y Aref (2012) utilizó el enfoque de reducción Rich Semantic Graphs (RSG) para generar los resúmenes abstractos de un solo documento. Redujeron el gráfico utilizando reglas heurísticas como sustitución, eliminación o fusión de nodos. Los nodos gráficos se obtienen mediante el uso de ontología de dominio, que consta de jerarquía de clases y relaciones. Bartakke et al. (2016) y Munot y Govilkar (2015) creó los gráficos semánticos para cada oración y luego usó la ontología para interconectar los conceptos para crear los subgráficos semánticos ricos y luego aplicó las reglas heurísticas para crear el gráfico semántico reducido final y luego lo introdujo en la herramienta de generación de lenguaje natural para crear el resumen resumen.
 - Gráficos AMR: representación de significado abstracto (AMR) Los gráficos son gráficos acíclicos etiquetados, arraigados y dirigidos de las oraciones. Los gráficos AMR proporcionan la representación semántica de la oración. Los nodos de los gráficos representan conceptos y los bordes representan la relación entre los conceptos. Los gráficos AMR representan una gran cantidad de información sobre el texto, como las entidades nombradas, el etiquetado de roles semánticos, la estructura de predicado-argumento y la relación de co-referencia (Foland y Martín, 2017). Los conceptos son en su mayoría palabras en inglés o PropBank⁸ marcos Aunque hay una serie de analizadores disponibles para analizar, pero en su mayoría analizadores JAMR⁹ se utiliza para realizar el análisis; realiza el análisis en dos pasos: primero identificando los conceptos clave utilizando el modelo semi-Markov, segundo identificando las relaciones entre los conceptos buscando el gráfico de expansión máximo conectado.
- Al generar el gráfico de origen para la oración, los conceptos similares se fusionan, al final de la fusión, una entidad se representa en el gráfico solo una vez. Así, independientemente

⁸<https://propbank.github.io/>.

⁹<https://github.com/jflanigan/jamr>.

de número de veces, el concepto ha aparecido en una oración, se representa como un solo nodo. Y esto, cuando se aplica a oraciones múltiples, conduce al gráfico final sin redundancia (Liu, Flanagan, Thomson, Sadeh y Smith, 2015). Los gráficos AMR se utilizan en tres representaciones, a saber, en forma conjunta para encontrar la similitud entre más de 1 gráfico AMR, en notación PEN-MAN y como una estructura de datos de gráficos (Vietnam, Sinh, Minh y Satoh, 2017).

Vilca y Cabezedo (2017) usó el análisis AMR para cada oración para generar el gráfico conceptual e incorporó información a nivel de Discurso para identificar los conceptos. Usaron estos conceptos para generar los resúmenes en lenguaje natural usando SimpleNLG. Liu et al. (2015) usó la transformación de gráfico a gráfico para una generación de resumen independiente del dominio. Usaron Treebank para crear gráficos AMR y luego transformaron estos gráficos para generar el resumen. Liao, Lebanoff y Liu (2018) creó los gráficos AMR para cada oración y luego los fusionó para crear los resúmenes utilizando la realización superficial. Han utilizado JAMR Parser para identificar los conceptos y la relación entre los conceptos.

Con la creciente investigación de redes neuronales con fines de resumen, también se han aplicado con éxito a la tarea de análisis AMR. Konstas, Iyer, Yatskar, Choi y Zettlemoyer (2017) usó el modelo basado en LSTM bidireccional en el lado del codificador y un modelo LSTM apilado en el lado del decodificador para análisis AMR. Foland y Martín (2017) usó el modelo bidireccional de memoria a corto plazo (LSTM) para capturar el contexto de la secuencia pasada y futura. Viet et al. (2017) creó un modelo convolucional de secuencia a secuencia junto con la linealización de gráficos para el análisis sintáctico AMR que ha superado tanto el análisis convencional AMR como los enfoques de generación de oraciones. Usaron modelos convolucionales en lugar de modelos LSTM para reducir la dependencia, lo que luego ayudó a reducir el recorrido del gráfico.

- Basado en unidades semánticas básicas: los gráficos AMR se centran solo en la transformación de gráfico a gráfico y el módulo de generación de texto no está desarrollado para ello. Para lidiar con eso, Li (2015) han usado información semántica para crear resúmenes abstractos de múltiples documentos usando BSU (Unidad Semántica Básica), el elemento más básico del texto coherente. Han construido una red de enlaces semánticos BSU donde BSU indica actores y receptores y consta de enlaces semánticos, nodos semánticos y reglas de razonamiento. Los enlaces semánticos ayudan a capturar el contexto alrededor de los nodos. El texto se puede generar de manera eficiente utilizando estos métodos.

2.2.4. Método semántico multimodal

La mayor parte de la información en la web no es puramente textual. En su mayoría contiene imágenes, videos, etc. junto con el texto. Kasture et al. (2014) ha dividido el marco general para el resumen multimodal en tres fases, a saber, la construcción del modelo semántico de conceptos, una calificación de conceptos sobre la base de factores como la integridad y la generación de oraciones. El dólar (2011) utilizaron la densidad de información como métrica para calcular el contenido obtenido a partir de métodos de resumen multimodal. Para crear los resúmenes multimodales abstractivos, primero crearon el modelo semántico utilizando la representación del conocimiento, donde construyeron los conceptos y luego extrajeron la información semántica de la clase específica limitada de imágenes. Después de extraer la información semántica, identificaron los conceptos importantes utilizando la métrica de densidad de información. Utilizaron los métodos de redacción para generar los resúmenes. UzZaman, Bigham y Allen (2011) creó resúmenes multimodales de oraciones complejas encontrando las entidades y la idea principal detrás del texto y luego agregó la estructura a las imágenes. En este enfoque se utiliza la ontología para encontrar los conceptos, pero no se utiliza una ontología generada automáticamente en estos enfoques, lo que conduce a una pérdida de tiempo. Además, no hay auto-

método automático para evaluar estos resúmenes, lo que limita la eficiencia de estos métodos.

Tabla 4 enumera las ventajas y desventajas de los enfoques basados en la semántica en la página 16.

2.3. Aprendizaje profundo y enfoque basado en redes neuronales

El aprendizaje profundo es parte de los métodos basados en el aprendizaje automático e implica datos de capacitación y aprendizaje. Implica múltiples capas de procesamiento no lineal para extraer las características del texto. El aprendizaje puede ser supervisado o no supervisado. Se basan en redes neuronales artificiales. Deep Learning se ha aplicado con éxito a varias tareas de PNL. Compras y Blunsom (2017) mostró que los analizadores basados en redes neuronales recurrentes (RNN) han logrado un rendimiento de vanguardia en el análisis de dependencia y constitucional. Los modelos RNN ayudan a predecir relaciones complejas que los enfoques simples estructurados o basados en la semántica no pueden hacer solos.

2.3.1. Introducción al modelo codificador-decodificador

El codificador y el decodificador son las redes neuronales individuales y funcionan juntos como una red neuronal combinada. La tarea del codificador es comprender las secuencias de entrada, mientras que la tarea del decodificador es determinar las secuencias y dar la salida. El codificador convierte las palabras en la representación de vectores que ayuda a capturar el contexto. En su mayoría, la incrustación de palabras se usa para representar las palabras en los codificadores, pero muchos han usado el modelo de bolsa de palabras (Rush, Chopra y Weston, 2015) también. Los decodificadores ayudan a encontrar la siguiente palabra en el resumen sobre la base de las palabras anteriores. Cuando tanto la entrada como la salida están en forma de secuencia, como en el caso del resumen de texto, los problemas de aprendizaje también se denominan problemas de aprendizaje Seq2Seq. Los modelos Codificador-Decodificador ayudan a resolver estos problemas. El mecanismo de atención se utiliza principalmente en modelos de secuencia donde la información se extrae del codificador en función de las puntuaciones de atención y el decodificador utiliza esta información. La atención ayuda a saber qué parte de la información necesitamos enfocar en una marca de tiempo particular.

Algunas redes comunes utilizadas en modelos de codificador-decodificador para resolver el problema de resumen abstracto son: -

- Redes neuronales convolucionales (CNN): aquí el tamaño de entrada siempre es fijo. Cada entrada de la red es independiente de entradas anteriores y futuras.
- Redes neuronales recurrentes (RNN): en la mayoría de las aplicaciones prácticas, el tamaño de entrada generalmente no es fijo. Además, las entradas no son independientes, sino que dependen unas de otras. Además, el número de predicciones requeridas como salida tampoco es fijo. Estos también se denominan problemas de aprendizaje de secuencias. Se agregan conexiones recurrentes a las redes neuronales, lo que ayuda a capturar la dependencia de las entradas con entradas anteriores o futuras.
- Redes de memoria a corto plazo (LSTM): RNN lee la entrada de izquierda a derecha y actualiza el estado después de cada palabra. Pero, cuando llegamos al final, la información de las primeras palabras se pierde. Para habilitar el olvido, la lectura selectiva y la escritura selectiva, se utilizan LSTM. Incluye las puertas para soportar las operaciones de olvidar, leer y escribir. Las unidades recurrentes cerradas (GRU) también se utilizan muy popularmente en el proceso de resumen abstracto. GRU son las variantes de LSTM, donde en lugar de diferentes puertas de entrada y de olvido, solo se usa una puerta. Forget gate no se usa explícitamente.

2.3.2. Trabajos recientes sobre resumen abstracto utilizando modelos de aprendizaje profundo

Baralis et al. (2013) utilizó redes neuronales para analizar los gráficos semánticos para generar resúmenes abstractos mediante un análisis lingüístico profundo. Utilizaron Minimal Recursion Semantics (MRS) para la representación semántica de gramáticas. Realizaron la desambiguación utilizando el modelo de máxima entropía. Ellos desarrollaron este

modelo para resolver el problema de alineación de los gráficos AMR. MRS se puede utilizar tanto para el análisis como para la generación de texto. [Niu et al. \(2017\)](#) propuso un enfoque de resumen abstracto de varios documentos mediante el uso de gráficos de fragmentos y redes neuronales. Utilizaron un modelo de lenguaje de red neuronal recurrente que ayudó a evaluar la calidad lingüística de la oración, lo que ayudó aún más a crear buenos resúmenes abstractos legibles.

El modelo simple de secuencia a secuencia asigna la secuencia de entrada a la secuencia de salida. [Jobson y Gutiérrez \(2018\)](#) utilizó el codificador-decodificador RNN junto con LSTM para crear los resúmenes. Usaron la incrustación de palabras con fines de capacitación y la función de atención para crear el vector de contexto en cada paso de tiempo. [Nallapati, Zhou, dos Santos, Gulehre y Lapata \(2016\)](#) usó el modelo de atención junto con el RNN para manejar los problemas de modelado de palabras clave y captura de la estructura jerárquica entre la oración y la palabra. Utilizaron el codificador bidireccional con GRU (Unidad recurrente Gated, utilizada para resolver los problemas de gradiente de fuga y explosión) -RNN y el decodificador unidireccional con GRU-RNN. Usaron el modelo de atención en los estados ocultos de la fuente y la capa softmax en el objetivo. [Rush et al. \(2015\)](#) usó la red neuronal de avance para trabajar en el resumen de texto a nivel de oración. Utilizaron el codificador basado en la atención y el decodificador basado en la búsqueda de haz para resumir el nivel de la oración. [Chopra, Auli y Rush \(2016\)](#) usó el modelo codificador-decodificador junto con el RNN condicional para resolver el problema similar a [Rush et al. \(2015\)](#). [Rossiello, Basile, Semeraro, Ciano y Grasso \(2016\)](#) usó las redes neuronales junto con RNN y modelos probabilísticos para crear resúmenes abstractos gramaticalmente correctos. Utilizaron el conocimiento previo junto con las redes neuronales para modelar el problema. [Liy, Lamy, Bingz y Wangy \(2017\)](#) usó el modelo codificador-decodificador de secuencia a secuencia para generar los resúmenes abstractos. Consideraron la información estructurada latente del texto para mejorar la calidad de los resúmenes. Usaron el decodificador generativo recurrente para traducir el código fuente a estados ocultos y luego regresar a las secuencias de palabras originales para generar el resumen.

[Canción, Huang y Ruan \(2018\)](#) propusieron un enfoque basado en el aprendizaje profundo llamado modelo codificador-descodificador de memoria a largo plazo a corto plazo en el que, en lugar de palabras, han utilizado frases como entrada para generar resúmenes abstractos. [Fan, Grangler y Auli \(2018\)](#) creó un enfoque de resumen abstracto controlable personalizado utilizando un modelo de redes neuronales convolucionales basado en codificador-decodificador de secuencia a secuencia. Crearon el resumen de acuerdo a las preferencias del usuario como entidad, cuya información quieren saber, el tamaño del resumen, parte del texto cuyo resumen quieren obtener.

La mayoría de los modelos basados en aprendizaje profundo se aplican a los sistemas genéricos de un solo documento. [Baumel, Eyal y Elhadad \(2018\)](#) creó los resúmenes abstractos específicos de la consulta de varios documentos utilizando el modelo basado en la atención sensible a la relevancia. [Nema, Khapra, Laha y Ravindran \(2018\)](#) utilizó el modelo de atención impulsada por la diversidad para resolver el mismo problema. [Nema et al. \(2018\)](#) han intentado resolver el problema de la generación de frases repetidas mediante un modelo de secuencia a secuencia utilizando un modelo de atención basado en la diversidad, que ayudó a lograr una ganancia del 28 % en la puntuación ROUGE-L.

Aunque el aprendizaje profundo se ha aplicado con éxito y ha surgido como uno de los enfoques prometedores para crear resúmenes abstractos. Pero la disponibilidad de un buen corpus grande para fines de formación sigue siendo un desafío. Además, la mayoría de los corpus son antiguos y, por lo tanto, no contienen las características morfológicas, semánticas y sintácticas actualizadas. No solo esto, sino que la mayoría de los corpus están solo en inglés. [Modaresi y Conrado \(2016\)](#) han creado un enfoque para crear un solo corpus de documentos para abordar los problemas mencionados anteriormente. [Lin, Sun, Ma y Su \(2018\)](#) usó el modelo Seq2Seq junto con el mecanismo de atención para resolver

tabla 1

Resumen de los resultados de la evaluación sobre el conjunto de datos DUC.

Técnica	Rango de puntuación de ROUGE-1
Basado en árboles	0,3–0,4
Basado en plantillas	0,21–0,35 (Oya, 2014)
Basado en Ontología	0,29–0,319 (Hennig, Umbrath y Wetzker, 2008b)
Basado en gráficos de frases principales y corporales	0,3 (máximo: 0,28) (Canción, Huang et al., 2018)
Basado en grafos semánticos	0,3–0,4 (0,417 como mejor puntuación) (Khan et al., 2018)
Basado en predicado-argumento	0,3–0,4
Basado en discurso multimodal	0,2–0,35 (Cohan et al., 2018)
Aprendizaje profundo	0,05–0,3
	0,28–0,47 (Joshi, Fidalgo y Alegre, 2018)

el problema de las repeticiones con los modelos Seq2Seq. Utilizaron unidades de compuerta convolucional junto con la codificación global en el lado del codificador y LSTM unidireccional en el lado del decodificador para realizar el resumen abstracto.

2.4. Enfoque basado en el discurso y la estructura retórica

Discourse Structure ayuda a capturar la estructura del texto y, por lo tanto, ayuda a encontrar las oraciones más importantes del texto ([Khan et al., 2018](#)). Aquí el objetivo principal es captar la representación interna del texto y convertirlas en las relaciones del discurso ([Hipola et al., 2014](#)). La teoría de la estructura retórica asume que los documentos pueden representarse en forma de árboles jerárquicos. A partir del discurso se crea una estructura retórica en forma de árbol. Se crea un algoritmo para asignar los pesos a estos elementos. Cuanto más arriba está el elemento en la estructura retórica, más alto se le da el peso. Luego, sobre la base de la longitud deseada para el resumen, se extraen los elementos. Los conectores discursivos ayudan a aumentar la coherencia y la cohesión del texto. El analizador de discurso se utiliza para crear los árboles de análisis y luego se crea la representación del árbol de discurso para las oraciones ([Gerani et al., 2014](#)). Estructura retórica ([Chengcheng, 2010](#)) es la forma de analizar el texto a nivel de cláusula y trata a nivel de documento. La relación retórica es la relación entre los dos textos que no se superponen. Se crea una estructura similar a un árbol para representar la coherencia en el texto. En él participan unidades discursivas con un elemento denominado núcleo y otro satélite ([Goyal y Eisenstein, 2016](#)). Los pasos principales son identificar la frase del texto, la creación de árboles de estructura retórica, el procesamiento de árboles para encontrar las oraciones importantes y no importantes mediante la búsqueda de los elementos centrales y luego la creación del resumen. La ventaja de las estructuras retóricas es que ayudan a crear resúmenes completos, gramaticalmente correctos y legibles.

[Gerani et al. \(2014\)](#) utilizó el enfoque basado en el discurso para resumir las reseñas de productos. Han utilizado aspectos y sus relaciones estructuradas para generar los resúmenes abstractivos analizando la estructura del discurso y las relaciones. Su enfoque no requiere conocimiento del dominio. El árbol del discurso basado en aspectos se crea para cada revisión, luego se fusionan para crear un gráfico de relaciones retóricas agregadas (ARRG). El PageRank ponderado se aplica a ARRG, para crear el subgráfico final llamado Árbol de jerarquía de aspectos (AHT) para crear el resumen final mediante la aplicación de la planificación del texto y la realización de oraciones. En el paso Realización de oraciones, han utilizado plantillas junto con el NLG.

[Tabla 3](#) enumera las ventajas y desventajas de los enfoques basados en el aprendizaje profundo en la página 15.

Abajo [entabla 1](#), hemos encontrado el rango de puntajes ROUGE obtenidos por los métodos mencionados anteriormente en el conjunto de datos DUC 2001, utilizado para crear resúmenes abstractos.

Tabla 2

Ventajas y desventajas del enfoque basado en estructuras.

	Ventajas	Desventajas
Basado en plantillas Métodos	Debido al hecho de que los fragmentos están llenos de la información extraída por los sistemas de Extracción de Información, proporcionan resúmenes altamente coherentes e informativos. También se pueden utilizar para el resumen de varios documentos.	Carecen de la diversidad ya que son en su mayoría predefinido. Carecen cuando el sistema necesita considerar la similitud y las diferencias entre los documentos (Kan, 2014).
Basado en Ontología Métodos	Los métodos basados en ontologías pueden manejar la incertidumbres asociadas con el texto fácilmente.	Los métodos basados en ontologías necesitan una buena ontología o diccionario que en su mayoría es creado por un experto en el dominio, por lo tanto, es un enfoque que requiere mucho tiempo.
Métodos basados en árboles	El uso del generador de idiomas mejora la calidad de resúmenes Producen resúmenes fluidos y menos redundantes (Kan, 2014) Son buenos para revisiones semánticas de un lead frase (Rananavare y Reddy, 2017).	Los métodos basados en árboles no tienen en cuenta el contexto; así se pierden muchas frases importantes del texto.
Métodos de frase principal y cuerpo		Los métodos Lead y Body Phrase producen redundantes, resúmenes menos gramaticales ya veces incompletos. El error de análisis conduce a la sustitución de frases incorrectas, lo que da lugar a resúmenes incorrectos.
Métodos basados en reglas	Los resúmenes son de alta densidad de información.	Debido al hecho de que las reglas y los patrones son en su mayoría hechos a mano, es muy tedioso crear estos resúmenes (Kasture1 et al., 2014).

Tabla 3

Ventajas y desventajas del aprendizaje profundo y el enfoque basado en redes neuronales.

Ventajas	Desventajas
1. Las redes neuronales ayudan a capturar el rol sintáctico adecuado de cada palabra. 2. Las redes neuronales pueden ayudar a reducir los errores gramaticales. 3. Las redes neuronales profundas ayudan a capturar el nivel intrínseco y alto representación distribuida de datos. 4. Los modelos de aprendizaje profundo ayudan a capturar la estructura semántica y sintáctica del texto en conjunto, lo que no es posible con técnicas simples de extracción o abstracción (Canción, Huang et al., 2018). 5. Ayudan a conseguir mejores ROUGE-Scores.	Se requiere una gran formación para captar la buena representación de el texto. Se basan en la co-ocurrencia estadística de palabras, lo que conduce a errores semánticos y gramaticales (Rossiello et al., 2016). Los modelos de aprendizaje profundo funcionan principalmente a nivel de oración, lo que hace que efectivos para la comprensión de oraciones pero no mucho para el resumen del documento. La mayoría de las veces, hay un problema de alineación entre el datos originales y de entrenamiento, lo que genera problemas durante el resumen abstracto (Jobson & Gutiérrez, 2018). Aunque, estas técnicas crean resúmenes cortos pero carencia en términos de densidad de información (Baumel et al., 2018). Sufren del problema de las repeticiones al crear resúmenes de varias oraciones (Chen y Bansal, 2018) (Canción, Zhao et al., 2018). Además, sufren el problema de la codificación lenta cuando el documento a resumir es largo (Chen y Bansal, 2018).

Tabla 4

Ventajas y desventajas del enfoque basado en la semántica.

	Ventajas	Desventajas
Gráficos AMR	Ayuda a producir contenido coherente, menos redundante y oraciones gramaticalmente correctas.	1. Los gráficos AMR están muy influenciados por la sintaxis y, por lo tanto, no admiten el descubrimiento y la manipulación de conceptos. 2. Se generan diferentes gráficos para la misma oración cuando se cambian a voz activa o pasiva. 3. Los gráficos AMR también sufren el problema de escasez de datos. 4. Los gráficos AMR dependen en gran medida de PropBank para las relaciones y, por lo tanto, le agregan restricciones y limitaciones de PropBank. 5. La falta de anotaciones alineadas crea desafíos al analizar gráficos AMR.
Métodos basados en INIT	Los métodos basados en INIT producen resúmenes no redundantes e informativos.	Un solo INIT no representa la oración completa. Por lo tanto, muchos INIT deben fusionarse para crear una oración completa (Genest & Lapalme, 2011). Los métodos basados en INIT descartan mucha información importante al intentar crear resúmenes significativos y gramaticalmente correctos. Los análisis incorrectos también conducen a la generación de resúmenes de baja calidad lingüística. (Khan et al., 2018)
basado en multimodal Métodos	Los métodos multimodales producen excelentes resúmenes de cobertura.	Se requiere evaluación manual ya que el documento contiene tanto imágenes como texto.
Gráfico semántico enriquecido Métodos basados	Producen menos redundancia y gramaticalmente resúmenes correctos.	Rich Semantic Graphs (RSG) se limitan principalmente a resumen de un solo documento.

3. Generación de resúmenes abstractivos

En esta sección, nos hemos centrado en los pasos que son la **parte importante de generar el texto**. A partir de nuestra encuesta, encontramos que la **compresión de oraciones, la fusión de conceptos, el cálculo de puntajes de ruta y la generación de resúmenes** son algunas partes comunes de un sistema de resúmenes abstractos. En esta sección, nos hemos centrado en las **técnicas utilizadas por varios investigadores para realizar estas tareas particulares**.

3.1. Compresión de oraciones

La compresión de oraciones consiste en **reducir la longitud de la oración conservando el significado original de la oración** (Nápoles, Callison-Burch, Ganitkevitch y Durme, 2011). Muchas veces, los **detalles irrelevantes** también están presentes en la oración y es importante **eliminarlos** para evitar el problema del espacio. Es uno de los temas más importantes de la comunidad de PNL y juega un **papel muy importante en la creación de resúmenes abstractos**. Según el tipo de estructura utilizada para la compresión de oraciones, estos métodos se dividen en métodos basados en **árboles, basados en discursos y basados en oraciones**. Enfoques basados en **árboles** (Galera y McKeown, 2007; Caballero y Marcu, 2000) crean las **oraciones compuestas** editando los **árboles sintácticos obtenidos al analizar el texto**. Pero **dependen en gran medida del analizador**, lo que los limita. Mientras que los enfoques **basados en oraciones** (McDonald, 2006) **crean directamente las oraciones pero carecen de lexicalización**. Enfoques basados en el discurso (Clarke y Lapata, 2008) **consideran la información discursiva del texto circundante para comprimir las oraciones**.

Según el **tipo de aprendizaje utilizado** para la compresión, también se dividen en **modelos supervisados y no supervisados**. Los enfoques **supervisados** involucran el **entrenamiento para encontrar las frases para la compresión**. Para encontrar las **probabilidades**, utilizan el **modelo generativo o el discriminativo**. En los modelos **generativos**, la **probabilidad de compresión** del objetivo se encuentra directa o indirectamente mediante el **uso de un canal ruidoso**. Mientras que, en modelos **discriminativos**, el objetivo principal es **reducir los errores de entrenamiento**. Principalmente **SVM, entropía máxima, árboles de decisión y aprendizaje de gran margen** se utilizan para el modelado **discriminativo** (Clarke y Lapata, 2008). Los enfoques **no supervisados** usan **reglas y modelos de lenguaje para comprimir la oración**.

En función de si se eliminan las palabras o se utilizan las nuevas frases o palabras, estos enfoques se dividen en **basados en eliminación y basados en generación** (Yu, Zhang, Huang y Zhu, 2018). En el enfoque basado en **eliminar**, las **palabras sin importancia se eliminan de la oración y, al unir el resto del texto, se crea la oración final**. En los modelos basados en **generación**, se utilizan la **inserción, la sustitución y el reemplazo**. Los modelos basados en generar **requieren una comprensión más profunda del texto**.

Caballero y Marcu (2000) han descrito el **enfoque probabilístico para comprimir las oraciones**. Han utilizado **métodos generativos basados en canales ruidosos y árboles de decisión** para comprimir la oración. Galera y McKeown (2007) también usó los **árboles sintácticos para comprimir la oración usando gramáticas libres de contexto síncronas**. McDonald (2006) usó el **aprendizaje automático**, el enfoque de aprendizaje en línea discriminativo (algoritmo relajado con infusión de margen, MIRA) para **aprender los pesos de las características y encontrar el texto que debe considerarse para la compresión**. Para crear los vectores de funciones y la oración final, utilizaron el **análisis sintáctico profundo de la oración al analizar sus oraciones con el analizador de dependencia y el analizador de estructura de frase**.

Nguyen y Ho (2004) usó **modelos ocultos de Markov** para comprimir la oración. CLASSY (Clustering, Linguistics and Statistics para el sistema de resumen) (Conroy, Schlesinger y Stewart, 2005), comprime las oraciones usando **heurísticas sintácticas y léxicas como eliminación de frases, identificación de entidades nombradas, modelos ocultos de Markov**. La **Programación Lineal Entera (ILP)** es también una de las fa-

enfoque más para la **compresión de oraciones**. No solo ayuda a **capturar la información estadística**, sino que también ayuda a **capturar la información a nivel de discurso**. La mayoría de estos enfoques **utilizan la información local para comprimir la oración**. Para resolver este problema, Clarke y Lapata (2008) usó el **aprendizaje no supervisado junto con la optimización restringida**, llamada Programación lineal entera (ILP) para comprimir la oración. Usaron información del discurso junto con la compresión de oraciones simples. Sahooa, Bhoib y Balabantaray (2018) usó Support Vector Machine (SVM) para comprimir las oraciones analizando primero las oraciones de Stanford Parser y luego usó estas oraciones analizadas con fines de capacitación.

Zajic, Lin, Dorr y Schwartz (2006) desarrolló un enfoque llamado TRIMMER para comprimir las oraciones para crear resúmenes de múltiples documentos. Recortaron los constituyentes sintácticos de la oración hasta que alcanzaron el punto de umbral y luego seleccionaron las oraciones con la mejor cobertura del tema. Pero su aplicabilidad se limita solo al inglés y carece de términos para encontrar las oraciones apropiadas.

Mani et al. (2010) llevó a cabo un experimento para descubrir cómo funcionan los algoritmos de aprendizaje automático para decidir los predicados que se agregarán a los fragmentos de oraciones para combinarlos y crear una oración no extractiva; y descubrió que SVM funciona razonablemente bien en comparación con otros clasificadores. Nápoles et al. (2011) extrajo el conjunto de paráfrasis que minimiza la longitud de la oración. Zhang et al. (2016) usó Programación Lineal Entera para generar las oraciones para el resumen.

Los modelos de aprendizaje automático por sí solos no han alcanzado el estado del arte, pero cuando se usan con modelos como Redes Neuronales Recurrentes (RNN), han dado muy buenos resultados. La mayoría de las técnicas de compresión involucran el procesamiento de información sintáctica, Filippova, Alfonseca, Colmenares, Kaiser y Vinyals (2015) han propuesto un enfoque que utiliza modelos de memoria a corto plazo a largo plazo basados en RNN para comprimir la oración. Entrenaron al modelo utilizando un corpus de aproximadamente 2 millones de oraciones. Primero analizaron la oración, luego crearon los vectores de incrustación de 518 dimensiones de tamaño para la oración utilizando el árbol de dependencia obtenido mediante el análisis y luego decodificaron la secuencia utilizando el procedimiento de búsqueda de haz. Su sistema logró una puntuación de legibilidad de 4,5 sobre 5 y una puntuación informativa de 3,8 sobre 5. Yu et al. (2018) usó los modelos basados en eliminar y generar junto con el modelo Seq2Seq (RNN bidireccional y Unidades recurrentes cerradas, GRU) para comprimir la oración.

3.2. Generalización y fusión de conceptos

Aquí los diferentes conceptos que aparecen en el texto se reemplazan por un concepto que es la generalización de todos los conceptos. Ayuda a reducir el texto. Belkebir y Guessoum (2016) realizó la generalización del concepto utilizando corpus de WordNet. Lo hicieron encontrando las oraciones generalizables generando las rutas de hiperonimia de los conceptos en una oración y luego reduciendo el tamaño de las versiones generalizables. Sentence Fusion también se utiliza para crear los resúmenes. En la fusión de oraciones, las oraciones relacionadas se entregan al sistema como entrada, las estructuras de dependencia de las oraciones se alinean para encontrar la información común y luego, sobre la base de la alineación, se crea el árbol de fusión. Las alineaciones se pueden crear a nivel de palabra, de frase, de subcadena, etc., pero para fines de fusión, se considera mejor la alineación a nivel de árbol de dependencias. La alineación ayuda a comprender cómo se relacionan las palabras entre sí. Marsi y Krahmer, (2005).

Barzilay y McKeown (2005) usó los árboles de dependencia para la fusión de oraciones para crear un resumen fluido, gramaticalmente correcto y no redundante utilizando la propiedad de información común entre las oraciones. Ellos crearon el fu-

sion retículos y los linealizó utilizando el generador de lenguaje natural FUF/SURGE para crear las oraciones finales. Filippova y Strube (2008) creó los resúmenes abstractos de varios documentos para biografías alemanas utilizando el concepto de fusión de oraciones. Utilizaron la fusión y la poda de frases informativas junto con la alineación de la estructura de dependencia para crear los resúmenes. Mehdad et al. (2013) construyó un gráfico de vinculación para las oraciones para encontrar las oraciones más relevantes y luego usó los gráficos de palabras junto con la generalización y la agregación para combinar las oraciones para formar los resúmenes informativos. Belkebir y Guessoum (2016) usó la fusión de conceptos para crear la técnica de generación de texto a texto para crear los resúmenes abstractos. Pero ellos solos no pueden crear un buen resumen debido al hecho de que identificar los fragmentos comunes es un gran desafío y luego usar la red de fusión para combinar las oraciones para formar la oración gramaticalmente correcta nuevamente es un problema complejo. También el enfoque tiene el problema de la expresión-referencia. Konstantas et al. (2017) utilizó la programación lineal entera (ILP) para encontrar y combinar las frases informativas para obtener la solución óptima global.

3.2.1. parafraseando

Reemplazar las frases con paráfrasis más cortas ayuda a reducir la longitud del texto original (Barzilay y Lee, 2003b). La detección de paráfrasis y la búsqueda de vínculos entre las frases y las oraciones no es una tarea fácil. Cohn y Lapata (2013) usó los resúmenes a nivel de oración para crear el resumen mediante la creación de un sistema de paráfrasis de extremo a extremo para no solo adquirir las paráfrasis sino también usarlas para crear las nuevas cadenas. Utilizaron la gramática de sustitución de árboles sincrónicos para crear el árbol de análisis comprimido y luego los entrenaron de manera discriminativa. Issa, Damonte, Cohen, Yan y Chang (2018) usó el análisis AMR junto con el análisis semántico latente para detectar las paráfrasis. A través del análisis semántico latente, intentaron encontrar la similitud semántica a través de la representación distribucional. Pero la precisión de AMR Parser debido al problema de formalismo y anotación también limitó el sistema de detección de paráfrasis. Chen y Bansal (2018) usó la paráfrasis seleccionando primero las oraciones más destacadas y luego reescribiéndolas para crear el resumen abstracto. Operaban tanto a nivel de oración como a nivel de palabra.

3.3. Cálculo de puntajes de ruta

Gráficos semánticos que incluyen Rich Semantic Graph, Dense Semantic Graph, AMR, etc. y son una de las formas más populares de crear resúmenes abstractos. Los grafos semánticos se pueden obtener de dos maneras. Una forma es representando la oración en sujeto-objeto-verbo y otra es encontrando las relaciones de dependencia entre las oraciones. Para crear el gráfico de una manera posterior, se debe calcular la puntuación de la ruta de dependencia más corta (Joshi, Wang et al., 2018). Bhargava, Sharma y Sharma (2016) calculó las puntuaciones sobre la base de la redundancia de oraciones superpuestas y la longitud del camino calculando la intersección de la posición de las palabras en la oración. Mehdad et al. (2013) modelo de lenguaje utilizado, información sobre sustantivos y verbos presentes en las oraciones para encontrar los caminos que son más legibles y fluidos.

3.4. Generación de resumen

Principalmente, las herramientas NLG se utilizan para la generación de resúmenes, ya que aumentan la fluidez y disminuyen los errores gramaticales. NLG se ha utilizado para agregar nuevas estructuras de vocabulario y lenguaje a las oraciones (Lloret et al., 2013). Figura 2 presenta los módulos del proceso de generación del lenguaje natural. Básicamente implica: -

- Planificación de texto: este paso también se denomina determinación de contenido. Es decidir qué información se incluirá en el resumen. Es el paso previo en la mayoría de los sistemas para la creación de resúmenes. También se puede dividir en selección de contenido y estructuración de texto.
- Planificación de oraciones: es organizar el contenido de manera que se produzca la especificación de la oración. Los límites de la oración se especifican en esta fase. Aquí el objetivo es organizar el texto en subpárrafos. Se considera la relación entre el texto y, sobre la base de la relación, las oraciones se fusionan para crear párrafos intermedios. Este paso también se conoce como Agregación de oraciones. Es para dar estructura al resumen. Machine Learning ayuda a aprender una buena estructura de resumen (Genest & Lapalme, 2011). Esta fase se divide nuevamente en cuatro fases, a saber, análisis léxico, análisis del discurso, agregación y expresión de referencia.
 - Análisis léxico: aquí las unidades léxicas individuales se eligen en función de varios factores como la fluidez, la variabilidad, el lenguaje y el estilo formal del lenguaje. Aquí se encuentra la palabra y la frase correctas para sustituir al texto para expresar la información (Dohare y Karnick, 2018). Se obtienen sinónimos de sustantivos y verbos para generar las palabras objetivo. Aquí las palabras de dominio se clasifican en elementos léxicos. Implica encontrar palabras semánticamente similares, sinónimos u otras palabras taxonómicamente relacionadas.
 - Análisis del discurso: aquí las oraciones individuales se generan a partir de los sinónimos y frases que se encuentran durante la fase de análisis léxico.
 - Proceso de agregación: Consiste en decidir cómo se unirán las oraciones para formar los párrafos intermedios. Muchas veces se utilizan las reglas hechas a mano para encontrar la manera de unir las oraciones.
 - Proceso de expresión referencial: Aquí se sustituye el sujeto por los pronombres. Aquí se eligen las palabras apropiadas que permiten distinguir las entidades.
- Realización: Aquí el objetivo es convertir los párrafos intermedios así obtenidos de la fase de planificación de oraciones a los párrafos finales. Aquí se corrigen los párrafos considerando la coherencia sintáctica, la corrección gramatical y de puntuación. Implica transformaciones morfológicas y sintácticas.
- Evaluación: aquí los párrafos se clasifican y clasifican de acuerdo con varios factores, como la coherencia entre las oraciones, los sinónimos de palabras más utilizados, etc.

Para resumir los trabajos realizados en este campo de investigación, los hemos enumerado en forma tabular en [Tabla 5](#) en la página 26 que enumera el nombre del autor, la medida de evaluación utilizada, el conjunto de datos utilizado y el tipo de resumen abstracto creado.

4. Herramientas utilizadas y desarrolladas

De nuestra encuesta, hemos encontrado que la mayoría de los sistemas de resumen abstracto consisten en 3 pasos, a saber, preprocesamiento, inferencia y generación de lenguaje natural. El preprocesamiento del texto implica la creación de la representación del texto e incluye la identificación de las entidades nombradas, la resolución de la correferencia, la búsqueda del etiquetado de partes del discurso, la construcción de árboles de dependencia, la construcción de árboles semánticos, etc. Inferir el texto incluye aprender la representación de texto obtenido del paso de preprocesamiento. Este paso incluye principalmente la fusión, la eliminación, la aplicación de algún modelo de aprendizaje como redes neuronales, etc. Y el paso final incluye la generación de lenguaje natural donde se genera el resumen final gramaticalmente correcto para el texto. Sobre la base de los pasos involucrados en la creación de resúmenes abstractos,

Tabla 5
Resumen de obras.

Autor	Método	conjunto de datos	Medida de evaluación	Individual/Multi-Documento
(Belkebir y Guessoum, 2016)	Fusión de Conceptos y Generalización			Documento único
(Cohn y Lapata, 2013)	Enfoque de compresión de oraciones	Artículos del periódico de británico Cuerpo Nacional DUC 2002	Basado en humanos Evaluación (coeficiente de Spearman para la distribución de calificaciones) ROUGE-N	Documento único
(Bhargava et al., 2016)	Basado en estructura (basado en gráficos + basado en reglas)			Documento único
(Kurisinkel et al., 2017)	Basado en árboles	DUC 2004, DUC 2007, TAC 2011	Métrica ROUGE, evaluación basada en humanos	Multi-Documento
(Yousfi-Monod & Príncipe, 2008)	Compresión de oraciones basada en árboles	Conjunto de datos de opiniones	Tasa de compresión, evaluación humana	Documento único
(Baralis et al., 2013)	Basado en Ontología Yago	DUC 2004	Máximo Marginal Relevancia (MMR), Humano Evaluación	Documento único
(Barzilay & Mc Keown, 2005)	Basado en árboles	DUC 2002	Basado en humanos Evaluación, (índice de compresión, corrección gramatical) ROUGE-1, ROUGE-2, ROUGE-SU4 Evaluación basada en humanos	Multi-Documento
(Oya et al., 2014) (Gerani et al., 2014)	Basado en plantillas Estructura del discurso + Basado en plantillas	Cuerpo AMI Reseñas de Amazon.com y Cnet.com		Documento único Multi-Documento
(Banerjee et al., 2017)	Basado en gráficos + ILP	DUC 2004 y DUC 2005	ROUGE-2, ROUGE-L, ROUGE-SU4 y evaluación humana	Multi-Documento
(Genest & Lapalme, 2011)	Basado en INIT		Puntaje piramidal, calidad lingüística y puntaje general de capacidad de respuesta ROUGE-1, ROUGE-2, Promedio	Multi-Documento
(Sahooa et al., 2018)	Basado en reglas	DUC 2002	Precisión, recuperación promedio y puntuación F promedio	Documento único
(Khan et al., 2018)	Resumen basado en gráficos semánticos	DUC 2002	Puntuaciones ROUGE-1, ROUGE-2 y Pyramid	Multi-Documento
(Bartakke et al., 2016)	Resumen basado en gráficos semánticos		Precisión y recuperación	Multi-Documento
(Moawad y Aref, 2012)	Resumen basado en gráficos semánticos enriquecidos	Estudiante graduado texto	tasa de compresión	Documento único
(Liu et al., 2015) (Le & Le, 2013)	Resumen basado en gráficos AMR Resumen basado en reglas sintácticas y discursivas	Banco AMR vietnamita NoticiasPaperSS	ROJO-1 ROJO-1 y ROJO-2	Documento único Documento único
(Ganesan et al., 2010)	Resumen abstracto basado en gráficos	Conjunto de datos de opinión	ROJO-1, ROJO-2, ROJO-SU4	Documento único
(Zhang et al., 2016)	Abstracción basada en estructura de argumento de predicado multilingüe resumen	DUC 2001	ROJO-1, ROJO-2, ROJO-SU4	Multi-Documento
(Vilca & Cabezudo, 2017)	Análisis Semántico y Basado en el Discurso	DUC 2002	Medida F-1, ROJO-1	Documento único
(Alshaina et al., 2017)	Resumen abstracto usando estructura de argumento-predicado Resumen	DUC 2002	Puntaje piramidal	Multi-Documento
(Canción, Huang et al., 2018)	abstracto usando aprendizaje profundo	CNN y DailyMail	ROUGE-N	Documento único
(Niu et al., 2017)	Grafo + Redes Neuronales Basadas en Resumen Abstractivo Atención Codificador-Decoder	DUC2004	ROJO-2, ROJO-SU4	Multidocumento
(Nallapati et al., 2016)	Decodificador Redes Neuronales Recurrentes	DUC2003, DUC2004, CNN	ROJO-1, ROJO-2, ROJO-L	Documento único
(Rush et al., 2015)	Aprendizaje profundo basado en el modelo de atención	Cuerpo de correo DUC 2003, Gigapalabras, DUC 2004	ROJO-1, ROJO-2, ROJO-L	Documento único
(Chopra et al., 2016)	Aprendizaje profundo basado en RNN convolucional	Gigapalabras, DUC 2004	ROJO-1, ROJO-2, ROJO-L	Documento único
(Liy et al., 2017)	Basado en aprendizaje profundo	Gigapalabras, DUC 2004, LCSTS	F-Medidas de ROUGE-1, ROJO-2, ROJO-L, ROJO-SU4 ROJO-1, ROJO-2, ROJO-L	Documento único
(Baumel et al., 2018)	Modelo de secuencia a secuencia Aprendizaje profundo basado en la relevancia Modelo de atención sensible	Debatepedia, DUC 2005, 2006 y 2007		Multi-Documento Consulta específica
(Nema et al., 2018)	Aprendizaje profundo basado en el modelo de atención impulsada por la diversidad	basado en Debatepedia conjunto de datos	ROJO-1, ROJO-2 y ROJO-L	Multi-Documento Consulta específica
(Fan et al., 2018)	Modelo de red neuronal convolucional codificador-decodificador	CNN-Dailymail	ROJO-1, ROJO-2 y ROJO-L	Personalizado Controlable Documento único

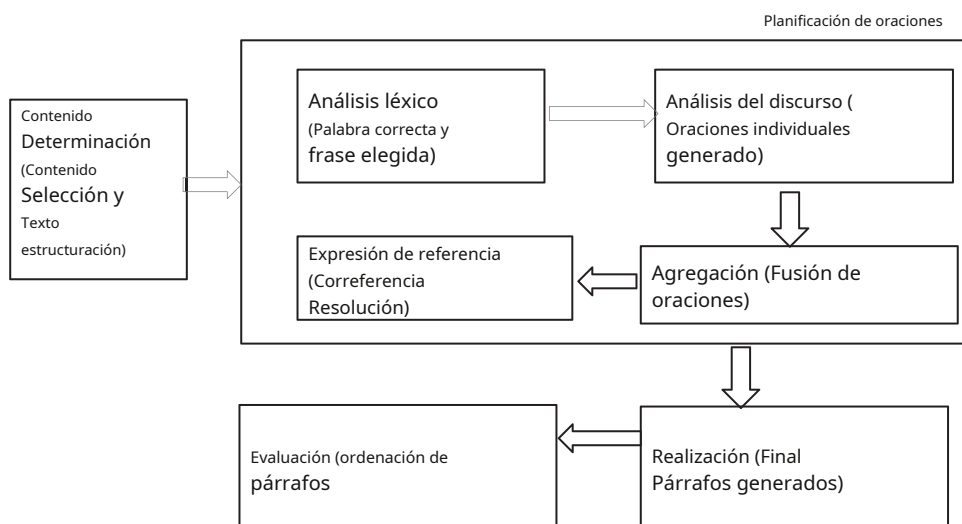


Figura 2. Pasos involucrados en la creación de resúmenes usando NLG.

técnicas de rización y herramientas de generación de lenguaje natural. A continuación se muestra la lista de herramientas junto con sus funcionalidades.

- Herramientas de preprocesamiento: estas son las herramientas que se utilizan para realizar el preprocesamiento durante el proceso de resumen y su objetivo es encontrar principalmente las entidades nombradas, realizar la segmentación de oraciones, el etiquetado de roles semánticos, crear los árboles de dependencia, ayudar en proceso de resolución de co-referencia, búsqueda de sinónimos, etc.
 - WordNet¹⁰: Es una base de datos léxica en inglés muy popular donde los sustantivos, verbos, adjetivos y adverbios se agrupan y organizan en synsets. Synsets se relacionan entre sí por sinónimos, hiperónimos, hipónimos, merónimos, holónimos, etc. y ayuda a encontrar las relaciones semánticas entre las palabras. Por lo tanto, wordnet ayuda en la clasificación del texto y encuentra los conceptos en el texto que luego ayuda en el proceso de resumen del texto.
 - FrameNet¹¹: Es una base de datos léxica en inglés legible por humanos y máquinas con ejemplos de anotaciones. Ayuda a asignar los roles semánticos y ayuda a encontrar las relaciones entre los predicados.
 - más disperso¹²: Es un analizador de gráficos basado en la estructura de frases. Tiene una gramática semántica extensa y extensible. El dólar (2011) usó el SParser para generar los resúmenes abstractos multimodales. ha sido usado en Greenbacker, McCoy, Carberry y McDonald (2011) realizar el análisis lingüístico identificando cada parte del texto en términos de tema y obtener los conceptos en términos de referentes parcialmente saturados.
 - Analizador MINIPAR¹³: Genest y Lapalme (2011) usó este analizador para encontrar los tripletes sujeto-verbo-objeto. Toma una oración como entrada, analiza la oración y encuentra las relaciones de dependencia entre las palabras de la oración.
 - Motor de extracción de información GATE¹⁴: GATE es una herramienta NLP basada en componentes y de código abierto. Ayuda en el etiquetado de POS y el etiquetado semántico. Ayudan a identificar los léxicos, entidades y ontologías. Tienen una capacidad de transducción de estado finito que los hace adecuados para

resumen como ayuda en la coincidencia de patrones y la identificación de anotaciones. También es posible usar reglas con ellos para un resumen efectivo. Genest y Lapalme (2011) lo utilizó para el análisis sintáctico.

- propank¹⁵: Es un corpus de texto anotado con información sobre proposiciones semánticas. Las relaciones predicado-argumento también están disponibles para Penn TreeBank sintáctico. Vilca y Cabezedo (2017) lo usó para el modelado de roles semánticos que ayuda a identificar preguntas como "Por qué", "Quién", "Dónde", "Cómo", etc.
- PNL de Stanford¹⁶: Ayuda a realizar muchas aplicaciones basadas en NLP como división de oraciones, etiquetado de parte del discurso, reconocimiento de entidades nombradas, análisis de distritos electorales, análisis de dependencias, extracción de información abierta, etc. Por lo tanto, ayuda a realizar resúmenes de texto. En la mayoría de los artículos, para crear los árboles de dependencia, se ha utilizado el analizador stanford NLP. Stanford CoreNLP ayuda a realizar la resolución de co-referencia.
- lingsoft¹⁷: Es una herramienta de análisis morfológico y sintáctico. Proporcionan formas básicas y gramaticales de palabras, que se pueden usar para etiquetar y organizar el resultado. Moawad y Aref (2012) usó Lingsoft para analizar sintácticamente la oración.
- Etiquetador de roles de SENNA: Khan et al. (2018) usó este analizador para determinar la estructura del argumento del predicado para la oración del texto. Junto con el modelado de roles, también ayuda a realizar el etiquetado de parte del discurso, el reconocimiento de entidades nombradas, la fragmentación y el análisis sintáctico. Crea marcos para cada verbo en la oración.
- LibSVM¹⁸: es un software para la clasificación de vectores de soporte. Ayuda a realizar máquinas de vectores de soporte ponderadas de clasificación multiclase para datos no balanceados. Support Vector Machines (SVM) ayuda a encontrar oraciones importantes mediante la categorización de texto, la fragmentación y el análisis de dependencia.
- Chunker de Illinios¹⁹: es una herramienta de software que divide el texto en palabras relacionadas semánticamente. Se ha usado (Oya et al., 2014) para identificar todas las frases nominales junto con

¹⁰ <https://wordnet.princeton.edu/>. <https://framenet.icsi.berkeley.edu/fndrupal/>

¹¹ [about. https://github.com/charlieg/Spaser](https://github.com/charlieg/Spaser). <https://gate.ac.uk/releases/>

¹² gate-7.0-build4195-ALL/doc/tao/splitch17.html. <https://gate.ac.uk/sale/tao/>

¹³ [splitch1.html](https://gate.ac.uk/sale/tao/splitch1.html).

¹⁴

¹⁵ <https://propank.github.io/>. <https://stanfordnlp.github.io/CoreNLP/simple.html>. <http://www.lingsoft.fi/en/499>. <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>. https://cogcomp.org/page/software_view/Chunker.

¹⁷ www.lingsoft.fi/en/499. <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>. https://cogcomp.org/page/software_view/Chunker.

¹⁸ cogcomp.org/page/software_view/Chunker.

¹⁹ [software_view/Chunker](https://cogcomp.org/page/software_view/Chunker).

el etiquetado de partes del discurso en la oración para crear resúmenes basados en plantillas.

- **VerbNet²⁰**: Es un léxico de verbos en inglés en línea, jerárquico e independiente del dominio. Asigna verbos a otros léxicos como WordNet. Los roles también están asociados con los verbos. Proporciona una descripción sintáctica de marcos y predicados semánticos para los verbos, lo que ayuda a restringir los roles temáticos. Se ha utilizado ampliamente para la creación de gráficos semánticos, cálculo de similitud de oraciones y eliminación de ambigüedades a nivel de token para resúmenes de texto.
- **Practica PNL²¹**: Son bibliotecas de Python que se construyen sobre SENNA y el extractor de dependencias de Stanford. Se ha utilizado para extraer la estructura predicado-argumento del texto (Alshaina et al., 2017).
- **Kit de herramientas NLTK²²**: NLTK ayuda a eliminar las palabras vacías, fragmentar la oración, encontrar la frecuencia de cada palabra. **katja** (2010) usó el kit de herramientas NLTK para etiquetar los fragmentos en las oraciones.
- **Kit de herramientas Stanford Core-NLP²³**: Ha sido ampliamente utilizado para realizar la resolución de co-referencia (katja, 2010).
- **JAMR Parser**: Se utilizan para AMR Parsing. Junto con la predicción del gráfico para cada oración, también proporciona la alineación entre la extensión de las palabras y los fragmentos del gráfico predicho, lo que ayuda durante la fase de generación del texto (Liu et al., 2015).

2. Herramientas de resumen

- **VISIÓN** (Demir et al., 2010): Es una herramienta para generar resúmenes textuales de gráficos de información. Transmite el mensaje subyacente junto con los aspectos más destacados mediante el uso de características visuales. Se ha utilizado su módulo de extracción visual (Greenbacker et al., 2011) para analizar la imagen y se ha utilizado el módulo de reconocimiento de intenciones para identificar las señales comunicativas presentes en los gráficos. Se utiliza principalmente para el resumen de documentos multimodales.
- **CONVOCATORIA** (Dohare y Karnick, 2018): Herramienta que crea resúmenes basados en plantillas. Es uno de los primeros sistemas de resumen multidocumento. Tiene dos componentes principales, principalmente planificador de contenido y generador lingüístico. El planificador de contenido determina la información importante que se incluirá en el resumen al combinar las plantillas de entrada, mientras que el generador lingüístico selecciona las palabras correctas que se utilizarán para crear un resumen coherente y gramaticalmente correcto. Pero las plantillas codificadas y la especificidad del dominio limitan su uso (Bhartiya y Singh, 2018).
- **RESUMEN²⁴**: Es un híbrido de Identificación de Temas, Interpretación y Generación de Lenguaje Natural. El sistema utiliza un diccionario y un tesoro SENSUS para identificar los temas y generalizarlos. La fusión de temas se realiza para la interpretación y luego se usa el generador de plantillas de frases para crear los resúmenes abstractos finales.
- **COMPENDIO** (Lloret et al., 2013) utiliza técnicas extractivas y abstractivas para crear el resumen final. El sistema primero identifica las oraciones relevantes sobre la base del análisis lingüístico superficial, la detección de redundancia mediante el uso de la herramienta de vinculación textual, la identificación del tema mediante el uso de TF-IDF y la detección de relevancia mediante el principio de cantidad de código; luego se realiza la comprensión y fusión de la información para generar los resúmenes abstractivos. Obtuvieron los gráficos de palabras para las oraciones, filtraron las rutas incorrectas en los gráficos de palabras y luego crearon los resúmenes.

- **MultiGEN²⁵**: Es una herramienta para realizar resúmenes de varios documentos en función de las similitudes y diferencias entre los documentos. Trabajan sobre la idea central de extracción temática para identificar las oraciones similares y reformulación a partir de temas comunes para generar el resumen.
- **Sistema MAR** (Carenini, Ng y Pauls, 2006): Summarizer of Evaluative Arguments, se utiliza para generar los argumentos evaluativos. Crea los resúmenes abstractos calculando la agregación de la información extraída y luego aplicándole la generación de lenguaje natural.
- **NAMAS²⁶**: Es un sistema de resumen de texto basado en abstracción neuronal desarrollado por Facebook, que utiliza Gigaword Corpus. Entrena al sistema y luego lo evalúa usando la puntuación ROUGE.
- **GISTEXTER** (Harabagiu et al., 2001): herramienta que crea resúmenes abstractos de un solo documento y de varios documentos. Requiere conocimiento del dominio. Crea resúmenes basados en plantillas extrayendo información utilizando el modelo de temas, CICERO. Debido a la extracción manual y las reglas lingüísticas, se requiere más tiempo y esfuerzo para generar resúmenes. También utiliza frases clave para lograr la cohesión en el resumen.
- **SemanticSumm**: Es una semántica²⁷ sistema de resumen basado en enfoque que utiliza gráficos AMR para crear resúmenes.

3. Herramientas de generación de lenguaje natural

- **SimpleNLG²⁸** para generar las oraciones (Genest & Lapalme, 2011).
- **Generador de lenguaje FUF o SURGE²⁹**: Se utiliza para generar oraciones fusionando y fusionando las frases. FUF es una herramienta de generación de lenguaje natural que se basa en la unificación de gramáticas. SURGE es un conjunto completo de gramática para FUF.

5. Desafíos y dirección futura

La complejidad del procesamiento del lenguaje natural plantea muchos desafíos para el resumen de texto abstracto. Hay muchos problemas de investigación abiertos en este campo que aún no se han resuelto. Algunos de los desafíos en este campo se enumeran a continuación:

- **Necesidad de medidas cuantitativas**: En función de si la evaluación se realiza comparando los resultados obtenidos mediante algún método automático ejecutado por el sistema o comparando los resultados obtenidos mediante la evaluación de personas, las técnicas de evaluación se dividen en: cuantitativas y cualitativas. Y, sobre la base de si la calidad del resumen se evalúa por sí mismo o se calcula el impacto del resumen en otras tareas como la legibilidad para determinar su eficacia, se dividen en evaluación intrínseca y extrínseca.

En la evaluación cuantitativa, la informatividad de la oración se analiza sobre la base del contenido disponible. ROUGE es la famosa métrica empleada para la evaluación cuantitativa. Mientras que en el análisis cualitativo, la satisfacción del usuario se evalúa frente a los resúmenes generados. Algunos investigadores (Banerjee et al., 2017) han utilizado la evaluación humana junto con el puntaje ROUGE para encontrar la calidad general del resumen.

²⁰ <https://verbs.colorado.edu/verbnet/>. https://nltk.readthedocs.io/en/latest/_modules/nltk/tools.html. <https://www.nltk.org/>. <https://stanfordnlp.github.io/CoreNLP/>. <https://www.isi.edu/natural-language/projects/SUMMARIST.html>.

²¹ <https://stanfordnlp.github.io/CoreNLP/>. <https://www.isi.edu/natural-language/projects/SUMMARIST.html>.

²² <https://www.isi.edu/natural-language/projects/SUMMARIST.html>.

²³ <https://www.isi.edu/natural-language/projects/SUMMARIST.html>.

²⁴

²⁵ <http://www.cs.columbia.edu/diglib/sumDemo/multiGen/main.html>.

²⁶ <https://github.com/facebookarchive/NAMAS>. https://github.com/summarization/semantic_summ. <https://github.com/simplenlg/simplenlg>. <https://www.cs.cmu.edu/Groups/AI/areas/nlp/nlg/fuf/0.html>.

²⁷ <https://github.com/simplenlg/simplenlg>. <https://www.cs.cmu.edu/Groups/AI/areas/nlp/nlg/fuf/0.html>.

²⁸ <https://www.cs.cmu.edu/Groups/AI/areas/nlp/nlg/fuf/0.html>.

²⁹

ROUGE es una métrica que se utiliza para realizar análisis cuantitativos y se basa en el recuerdo. Calcula el número de n-gramas superpuestos entre el resumen generado por el sistema y el resumen escrito por humanos. Hay muchas variantes de la métrica ROUGE como ROUGE-2, donde las secuencias de palabras de tamaño 2 se consideran para la comparación; ROUGE-L, donde se considera la subsecuencia común más larga para la comparación; ROUGE-SU4, donde se comparan los bigramas saltados con los unigramas. La puntuación de la pirámide también se ha utilizado con fines de evaluación, ya que puede evaluar la calidad más allá de la coincidencia a nivel de palabra (Li, 2015).

Banerjee et al. (2017) utilizó la informatividad y la calidad lingüística como 2 medidas para la evaluación basada en humanos. Oyá et al. (2014) utilizó un método humano de 5 escalas basado en la fluidez, demasiados errores gramaticales, información y cobertura para encontrar la calidad general del resumen. Cohn y Lapata (2013) utilizaron el enfoque de evaluación humana junto con el coeficiente de Spearman para comparar los resultados de su enfoque de compresión de oraciones abstractiva para garantizar la idoneidad de la distribución de las calificaciones que utilizaron. Baralis et al. (2013) utilizó la evaluación humana sobre la base de la legibilidad para encontrar la eficacia de su sistema. En la mayoría de los trabajos, la valoración de los resúmenes se ha realizado mediante ROUGE Scores. Los puntajes ROUGE ayudan a medir la cobertura, pero no ayudan a encontrar la coherencia y otros factores como la no redundancia, ya que solo calculan la repetibilidad de los N-gramas (Canción, Huang et al., 2018). También se ha observado que se ha obtenido la misma puntuación ROUGE para diferentes resúmenes de un mismo texto cuando sus contenidos son completamente diferentes y cuando la superposición también es casi diferente (Mehta, 2016). Incluso los puntajes de ROUGE no brindan información para ayudar a especificar las fortalezas y debilidades del resumen (Carenini y Cheung, 2008). Más trabajo para encontrar las variantes de ROUGE Score es la necesidad del día (Yao et al., 2017). Baume et al. (2018) y Zhang et al. (2016) han abordado el tema del diseño de medidas cuantitativas que pueden encontrar la calidad de los resúmenes abstractivos. ROUGE Score funciona bien para resúmenes extractivos, pero no es una buena métrica para evaluar resúmenes abstractivos, ya que necesitan una métrica que pueda encontrar la superposición semántica que la superposición de palabras.

Pyramid Score es una de las métricas utilizadas para evaluar el resumen desde la perspectiva semántica. Es un método de puntuación basado en anotaciones en el que las unidades de contenido de resumen (SCU) se obtienen de los resúmenes del modelo y luego se asigna el peso a cada SCU en función de la frecuencia de aparición en los resúmenes de referencia humana. Pero, se requiere mucha intervención humana para realizar esta evaluación para evaluar el contenido semántico del texto. AutoPyramid es una extensión del enfoque Pyramid donde Passonneau, Chen, Guo y Perin (2013) trató de reducir este esfuerzo manual al automatizar el proceso de encontrar si la SCU está presente en el resumen generado por el sistema o no. Utilizaron la programación dinámica y el método del vector latente para resolver este problema. Pero no pudieron cuantificar la calidad del resumen sobre la base del acuerdo y la contradicción con el resumen de referencia humana. Yang, Passonneau y de Melo (2016) creó un enfoque más llamado PEAK (Pyramid Evaluation via Automated Knowledge Extraction) para evaluar automáticamente la SCU mediante el uso de algoritmos gráficos y de extracción de información. Pero este enfoque también carecía de muchas cosas, como que no tiene en cuenta la resolución de la co-referencia, la identificación de paráfrasis y no logró modelar la contradicción. Por lo tanto, para resolver los problemas mencionados anteriormente de los métodos Pyramid, ROUGE y Auto-pyramid para evaluar los resúmenes abstractivos; Vadapalli, Kurisinkel, Gupta y Varma (2017) creó una métrica, Similitud semántica para resumen abstracto (SSAS) para evaluar los resúmenes abstractos en

nivel de inferencia mántica. Utilizaron un análisis semántico profundo para encontrar las medidas léxicas y semánticas para calcular la similitud entre los resúmenes generados por el sistema y los generados por humanos. Pero en la actualidad, este enfoque lleva demasiado tiempo en comparación con los métodos ROUGE, Pyramid, Auto-pyramid y PEAK. Por lo tanto, se requiere más trabajo en paralelo y encontrar mejores formas de obtener los vectores de características para hacer que SSAS sea más efectivo.

- **Cuestión de aprendizaje de palabras raras en redes neuronales:** Canción, Zhao y Liu (2018) mencionó cómo las redes neuronales o el aprendizaje de secuencia a secuencia no logran preservar el significado de resumen. Las palabras raras a veces causan problemas en los modelos seq-2-seq debido al hecho de que las palabras raras generalmente ocurren menos en el sistema de entrenamiento y, por lo tanto, el sistema no aprende sus patrones. Además, porque en los enfoques de aprendizaje profundo, si se usan individualmente, la estructura sintáctica no se alimenta explícitamente, por lo que carecen de la estructura sintáctica y son parciales. Los autores también han mostrado algunos ejemplos en los que el modelo Seq2Seq no logra capturar el verbo principal de la oración. Para abordar este problema, se deben explorar más esfuerzos que combinen la estructura sintáctica del texto con las redes neuronales.
- **Disponibilidad de conjuntos de datos:** La disponibilidad de un cuerpo de formación bueno y amplio es uno de los principales obstáculos en esta área; la mayoría de los conjuntos de datos pertenecen a artículos de noticias. De la encuesta, hemos observado que la mayoría de los sistemas de resumen abstracto han utilizado DUC30y TAC31conjunto de datos El conjunto de datos de Gigawords y el conjunto de datos de correo diario de CNN también son conjuntos de datos famosos entre la comunidad de aprendizaje profundo. Se requiere más trabajo en la creación de buenos conjuntos de datos para la comunidad de investigación. Los modelos de aprendizaje profundo han demostrado ser uno de los modelos efectivos para modelar el problema de resumen abstracto. Pero la gran cantidad de datos necesarios para fines de capacitación es uno de los principales problemas con el uso de estos modelos. La mayoría de los modelos de aprendizaje profundo se aplican para el resumen extractivo de un solo documento y hay muy pocos trabajos disponibles sobre el resumen abstracto específico de consultas. El motivo es la disponibilidad de conjuntos de datos a gran escala necesarios para realizar resúmenes específicos de consultas. Nema et al. (2018) han creado un conjunto de datos32usando Debatepedia33por solucionar este problema. La alineación de los datos también es uno de los problemas para extraer oraciones de origen y de destino (Mehta, 2016). La mayoría de los datos están alineados a nivel de documento y no a nivel de oración. Fan et al. (2018) han creado los resúmenes abstractivos controlables y han permitido a los usuarios mencionar la parte del texto que quieren resumir. Pero para crear el resumen de la parte específica del texto, se enfrentaron al desafío del conjunto de datos. Tuvieron que crear su propio conjunto de datos ya que no había ningún conjunto de datos fácilmente disponible que pudiera usarse para realizar este tipo de resumen. Alinearon los resúmenes con los documentos completos.
- **Necesidad de buenos algoritmos para la fusión y generalización de conceptos.:** La fusión de oraciones es una de las áreas más desafiantes del campo de resumen abstracto. No hay mucho trabajo hecho en el campo de la generalización y fusión de conceptos. Belkebir y Guessoum (2016) en su artículo han utilizado la regla de los sustantivos y la regla de los adjetivos para la fusión de conceptos, pero se requiere más trabajo para encontrar más reglas para realizar la fusión de conceptos. Su enfoque generó el conjunto de oraciones generalizables, pero su algoritmo tiene una complejidad de espacio exponencial y, por lo tanto, se requiere más trabajo para encontrar algoritmos que puedan reducir el espacio de oraciones generalizables. Mehdad et al. (2013) abordó la necesidad de encontrar buenos enfoques para la detección de comunidades y la fusión de oraciones para textos informales como información de transcripciones de reuniones.

30<https://duc.nist.gov/>.

31<https://tac.nist.gov/>.

32<https://github.com/PrekshaNema25/DiversityBasedAttentionMechanism>.

33<http://www.debatepedia.org/en/index.php/>.

mación Las oraciones comunitarias son aquellas oraciones del texto que se pueden fusionar para crear una oración abstracta.

- **Problemas de escalabilidad:** Aunque ha habido mucha investigación en este campo, los algoritmos requieren muchos datos y potencia para dar buenos resultados con documentos largos (Singhal, Vats y Karnick, 2018). Además, la mayor parte del trabajo en este campo se realiza en oraciones simples y compuestas. Pero se requiere más trabajo para encontrar enfoques escalables que consideren las oraciones compuestas complejas (Sahoo et al., 2018). Incluso cuando las redes neuronales se aplican al resumen de un documento largo, sufre el problema de la codificación lenta e imprecisa debido al hecho de que el mecanismo de atención se aplica principalmente y analiza todas las palabras codificadas con fines de decodificación (Chen y Bansal, 2018).
- **Cálculo de similitud semántica:** El cálculo de la similitud semántica entre las oraciones es uno de los problemas del campo del procesamiento del lenguaje natural. Desempeña un papel muy importante en el proceso de resumen abstracto, ya que ayuda a encontrar los conceptos y los conceptos son el corazón de los sistemas de resumen abstracto. En su mayoría, la coocurrencia de palabras, la base de datos léxica y los resultados del motor de búsqueda se utilizan para calcular la similitud semántica entre las palabras o las oraciones. Los modelos basados en la co-ocurrencia de palabras calculan esta similitud comparando el vector de consulta y el vector del documento, pero como no consideran el orden y el contexto de las palabras, no pueden capturar la similitud semántica de manera muy eficiente. Las bases de datos léxicas como WordNet también se utilizan para calcular la similitud entre frases u oraciones, pero la coincidencia directa de palabras o frases con la información de la base de datos léxica y el hecho de que el significado de la palabra difiere de un corpus a otro las limita. Incluso, los métodos basados en motores de búsqueda no dan muy buenos resultados debido al hecho de que las palabras con significados opuestos también aparecen junto con los resultados del motor de búsqueda. Para resolver estos problemas (Pawar & Mago, 2018) crearon un enfoque al crear los vectores semánticos para las oraciones utilizando WordNet como fuente léxica. Pero también se requieren más esfuerzos para extender estos métodos a varios dominios y analizar cómo los resultados difieren al usar diferentes ontologías. La mayoría de los trabajos han utilizado WordNet para encontrar la información léxica en el texto para crear los gráficos, Mehdad et al. (2013) mencionó la necesidad de encontrar cómo cambian los gráficos con el uso de diferentes fuentes de conocimiento como Yago Ontology o DBPedia.
- **Necesidad de aumentar la eficiencia de los gráficos AMR:** AMR es un grafo enraizado, dirigido y acíclico que se utiliza para representar la información semántica de una oración del texto. Los gráficos AMR se basan en el conjunto de marcos PropBank, por lo que la limitación del conjunto de marcos limita los gráficos AMR. Cuando los gráficos AMR se utilizan con fines de resumen, los gráficos individuales se fusionan mediante la identificación de conceptos similares (Liu et al., 2015). Pero, a medida que aumenta el tamaño del texto, aumenta el número de gráficos AMR, la fusión conduce a las complejidades. Por lo tanto, se deben descubrir más algoritmos de decodificación como la relajación lagrangiana o algoritmos aproximados para hacer que el enfoque basado en AMR sea más efectivo. Liu et al. (2015) también han planteado la necesidad de realizar tanto la resolución de co-referencia de entidad como de evento para hacer que la fusión de gráficos sea más eficiente. Además, para identificar los bordes que se seleccionarán como candidatos para el resumen, se identifican los subgráficos, pero principalmente la predicción del subgráfico se encuentra en el nivel de la oración. Identificar cómo se pueden encontrar los subgráficos a nivel de documento puede ayudar a lograr una mayor cobertura. La construcción de gráficos AMR depende de los analizadores disponibles, lo que limita su eficiencia, ya que todo el concepto de resumen mediante gráficos AMR depende de la identificación del concepto. Y la eficiencia de la identificación de conceptos depende de la eficiencia de los analizadores. Por lo tanto, con la mejora de los modelos de análisis, los modelos de resumen basados en AMR Graphs pueden ser mejorados.
- **Necesidad de plataforma única para especificar las ontologías:** Hay una serie de ontologías externas disponibles para su uso. Existe la necesidad de una plataforma única que pueda acomodar todas las ontologías explícitas y, por lo tanto, ayudará a obtener el extenso sistema de resumen abstracto para diferentes dominios. Xiang, Jiang, Chang y Sui (2015) han tratado de abordar este problema mediante la creación de un enfoque de coincidencia de ontologías llamado ERSOM, que encuentra las entidades relacionadas semánticamente entre diferentes ontologías.
- **Necesidad de sistemas de resúmenes abstractivos basados en varios idiomas:** El resumen entre idiomas consiste en producir el resumen de un texto escrito en algún idioma de origen, como el sánscrito, en algún otro idioma de destino, como el inglés. En esta era de la información, no todos los documentos son del mismo idioma. Diferentes documentos son de diferentes idiomas. La creación del resumen en varios idiomas ayudará a los lectores desconocidos a conocer la esencia del documento. A partir de la encuesta, hemos descubierto que para esta tarea se han utilizado principalmente técnicas de resumen extractivo y no se ha trabajado mucho en la creación de resúmenes abstractos para sistemas multilingües. Algunas de las obras de Yao, Wan y Xiao (2015) y Zhang et al. (2016) han utilizado la comprensión basada en frases y los argumentos de predicados utilizando el proceso de traducción automática junto con la programación lineal entera, respectivamente, para la generación de oraciones de documentos en varios idiomas. Se requiere más trabajo en este campo ya que el lenguaje cruzado impone una complejidad adicional al resumen abstracto debido al hecho de que no solo se requiere concisión, información y coherencia, sino que la calidad de la traducción también es un aspecto importante que decide la calidad del resumen (Zhang et al., 2016).
- **Necesidad de sistemas de resumen abstracto específicos de consulta basados en aprendizaje profundo:** Los problemas específicos de la consulta son aquellos en los que los resúmenes resaltan puntos relevantes en el contexto de la consulta y, en su mayoría, se han aplicado técnicas extractivas para generar resúmenes específicos de la consulta, pero sufren el problema de coherencia y esta probabilidad de obtener un resumen incoherente es muy alta en caso de consulta. -Problema de resumen específico ya que sin conocer el contexto y simplemente conectar las oraciones, es difícil resolver el problema de la co-referencia. Baumel et al. (2018) usó el aprendizaje profundo como un enfoque para crear resúmenes abstractos específicos de consultas. Nema et al. (2018) planteó el problema de las frases repetidas mediante el uso de modelos basados en codificadores y decodificadores al intentar generar resúmenes abstractos específicos de consultas. La mayoría de los modelos abstractos basados en el aprendizaje profundo se aplican para el resumen genérico de un solo documento. Hay muy pocos trabajos disponibles sobre resúmenes de múltiples documentos específicos de consultas. Se requiere más trabajo para identificar los enfoques para abordar este resumen.

6. Conclusión

El resumen abstracto es un tema interesante de investigación entre la comunidad de PNL y ayuda a producir resúmenes coherentes, concisos, no redundantes y ricos en información. La idea del artículo es presentar los estudios recientes y los avances realizados en este campo para ayudar a los investigadores a familiarizarse con las técnicas actuales, los desafíos existentes y las sugerencias para el trabajo futuro en esta área. Junto a estos también hemos mencionado las herramientas que se han utilizado en diversas investigaciones relacionadas con la síntesis abstractiva. La evaluación de resúmenes es un gran desafío en este campo. El análisis semántico y el análisis del discurso, junto con las nuevas tecnologías emergentes, como las redes neuronales, ayudan a superar las dificultades asociadas con el resumen abstracto.

Declaración de contribución de autoría CRediT

Som Gupta: Conceptualización, Metodología, Validación, Visualización, Escrita tanto en versión original como revisada. **S. K. Gupta:** Ayudó a Finalizar el tema, Decidió la metodología, Supervisó durante el proceso de redacción, Validó los contenidos, Revisó los contenidos.

Referencias

- Alshaina, S., John, A. y Nath, AG (2017). Resumen abstracto de varios documentos ción basada en estructura de argumento de predicado. En *Conferencia internacional sobre procesamiento de señales, informática, comunicación y sistemas de energía (SPICES)* (págs. 32–37). hacer:10.1109/ESPECIAS.2017.8091339.
- Banerjee, S., Mitra, P. y Sugiyama, K. (2017). Resumen abstracto de varios documentos rización usando compresión de oraciones múltiples basada en ilp. En *Actas IJCAI5 de la 24ª conferencia internacional sobre inteligencia artificial* (págs. 1208–1214). Baralis, E., Cagliero, L., Jabeena, S., Fiori, A. y Shah, S. (2013). Resumen multi-documento marización basada en la ontología yago. *Sistemas expertos con aplicaciones*, 40(17), 6976–6984. hacer:10.1016/j.eswa.2013.06.047.
- Bartakke, D., Sawarkar, DS y Gulati, A. (2016). Un enfoque basado en la semántica para ab-resumen de texto de documentos múltiples stractive. *Revista internacional de investigación innovadora en ingeniería informática y de comunicaciones*, 4(7), 13620–13628. Barzilay, R. y Lee, L. (2003a). Aprender a parafrasear: una aplicación no supervisada proach utilizando la alineación de secuencias múltiples. *En Actas del HLT-NAACL 2003* (págs. 16 a 23).
- Barzilay, R. y Lee, L. (2003b). Aprender a parafrasear: un enfoque no supervisado utilizando la alineación de múltiples secuencias. En *Procedimientos de HLT-NAAC* (págs. 16–23). Barzilay, R. y McKeown, KR (2005). Fusión de oraciones para noticias multidocumento resumen. *Lingüística Computacional*, 31(3), 297–327.
- Baumel, T., Eyal, M. y Elhadad, M. (2018). *Resumen abstracto centrado en la consulta: Incorporación de relevancia de consulta, cobertura de múltiples documentos y restricciones de longitud de resumen en modelos seq2seq* (págs. 297–327). Biblioteca de la Universidad de Cornell.
- Belkebir, R. y Guessoum, A. (2016). Generalización y fusión de conceptos para la abstracción, generación de oraciones tivas. *Sistemas expertos y aplicaciones*, 53, 43–56. hacer:10.1016/j.eswa.2016.01.007.
- Bhargava, R., Sharma, Y. y Sharma, G. (2016). Atssi: resumen de texto abstractivo utilizando la infusión de sentimiento. En *Duodécima multiconferencia internacional sobre procesamiento de la información* (págs. 404–411). hacer:10.1016/j.procs.2016.06.088.
- Bhartiya, R. y Singh, A. (2014). Un enfoque semántico del resumen. *CoRR*. eprintarXiv:1406.1203
- Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. y Hellmann, S. (2009). Dbpedia: un punto de cristalización para la red de datos. *Revista de semántica web: primer vistazo*, 154–165. hacer:10.1016/j.websem.2009.07.002.
- Buyts, J. y Blunsom, P. (2017). *Análisis de gráficos semánticos neuronales incrementales robustos*. 40. Biblioteca de la Universidad de Cornell.
- Carenini, G. y Cheung, JCK (2008). Resumen extractivo vs. basado en nlg marización del texto evaluativo: el efecto de la controversia del corpus. En *INLG '08 actas de la quinta conferencia internacional de generación de lenguaje natural: 1* (págs. 33–41).
- Carenini, G., Ng, R. y Pauls, A. (2006). Resumen multi-documento de evaluación texto. En *11ª CEAC 2006: 61* (págs. 305–312).
- Carenini, G., Ng, RT y Pauls, A. (2012). Resumen multi-documento de evaluación tivo. 0 (0). <https://doi.org/10.1111/j.1467-8640.2012.00417.x>
- Chen, Y.-C. y Bansal, M. (2018). Resumen abstractivo rápido con refuerzo-se-reescritura de oraciones elegidas. *LCA*.
- Chengcheng, L. (2010). Resumen automático de texto basado en estructura retórica teoría. En *Conferencia internacional sobre modelado de sistemas y aplicaciones informáticas* (págs. 595–598).
- Chopra, S., Auli, M. y Rush, AM (2016). Resumen de oraciones abstractivas con redes neuronales recurrentes atentas. En *Conferencia: Actas de la conferencia de 2016 del capítulo norteamericano de la asociación de lingüística computacional: tecnologías del lenguaje humano*.
- Clarke, J. y Lapata, M. (2008). Inferencia global para la compresión de oraciones: una inferencia enfoque de programación lineal entera. *Revista de Investigación de Inteligencia Artificial*, 31, 399–421.
- Cohan, A., Derroncourt, F., Kim, DS, Bui, T., Kim, S., Chang, W. y Goharian, N. (2018). Un modelo de atención consciente del discurso para el resumen abstracto de documentos largos. arXiv:1804.05685v2[cs.CL]
- Cohn, T. y Lapata, M. (2013). Compresión de oraciones parafrásticas con un carácter métrica basada en ter: apriete sin eliminación. En *Actas del taller sobre generación de texto a texto monolingüe*.
- Conroy, JM, Schlesinger, JD y Stewart, JG (2005). Multi elegante basado en consultas resumen del documento. *Conferencia de comprensión de documentos de 2005*.
- Sunitha, C., Jaya, A. y Ganesh, DA (2016). Un estudio sobre resumen abstracto Técnicas de ción en lenguas indias. En *Cuarta conferencia internacional sobre tendencias recientes en informática e ingeniería* (págs. 25–31).
- Demir, S., Oliver, D., Schwartz, E., Elzer, S., Carberry, S. y McCoy, KF (2010). Enterrar- vista activa en gráficos de información. *W4A2010*.
- Dineshnath, G. y S. Saraswathi, D. (2017). Resumen de texto usando abstractive métodos. *Revista de Redes de Comunicaciones y Tecnologías Emergentes (JNCET)*, 7 (12), 26–29.
- Dineshnath, G. y S. Saraswathi, D. (2018). Encuesta completa para texto abstracto resumen. *Revista Internacional de Innovaciones y Avances en Ciencias de la Computación*, 7(1), 215–219.
- Dohare, S. y Karnick, H. (2018). Estudio del estado del arte en lenguaje natural generación: Tareas básicas, aplicaciones y evaluación. *Revista de Investigación de Inteligencia Artificial*, 61, 65–170. arXiv:1703.09902v4.
- Embar, VR, Deshpande, SR, Vaishnavi, AK, Jain, V. y Kallimani, J. (2013). sArAmsha - un resumen abstracto de Kannada. En *Avances en Computación, Comunicaciones e Informática (ICACCI), 2013 Conferencia Internacional sobre* (págs. 540–544). hacer:10.1109/ICACCI.2013.6637229.
- Fan, A., Grangler, D. y Auli, M. (2018). Resumen controlable abstractivo. En *Actas del segundo taller sobre traducción y generación de máquinas neuronales* (págs. 45–54).
- Filippova, K., Alfonseca, E., Colmenares, CA, Kaiser, L. y Vinyals, O. (2015). Sen-compresión de tenencia por borrado con lstms. En *Actas de la conferencia de 2015 sobre métodos empíricos en el procesamiento del lenguaje natural*.
- Filippova, K. y Strube, M. (2008). Fusión de oraciones a través de la compresión de gráficos de dependencia. En *Actas de la conferencia de 2008 sobre métodos empíricos en el procesamiento del lenguaje natural* (págs. 177–185).
- Foland, W. y Martín, JH (2017). Representación de significado abstracto analizándonos-ing lstm redes neuronales recurrentes. En *Actas de la 55.ª reunión anual de la asociación de lingüística computacional* (págs. 463–472). hacer:10.18653/v1/P17-1043.
- Galera, M. y McKeown, KR (2007). Gramáticas de markov lexicalizadas para oraciones compresión. En *Conferencia del capítulo norteamericano de la asociación de lingüística computacional*.
- Ganesan, K., Zhai, C. y Han, J. (2010). Opinosis: un enfoque basado en gráficos para la abstracción resumen tivo de opiniones altamente redundantes. En *23ª conferencia internacional de lingüística computacional* (págs. 340–348).
- Genest, PE y Lapalme, G. (2011). Marco para el uso de resúmenes abstractivos generación de texto a texto. En *Taller de generación de texto a texto monolingüe* (págs. 64–73).
- Gerani, S., Mehdad, Y., Carenini, G., Ng, RT y Nejat, B. (2014). suma abstractiva marización de reseñas de productos utilizando la estructura del discurso. En *Conferencia de 2014 sobre métodos empíricos en el procesamiento del lenguaje natural* (págs. 1602–1613).
- Goyal, N. y Eisenstein, J. (2016). Un modelo conjunto de estructura del discurso retórico y resumen. En *Actas del taller sobre predicción estructurada para PNL* (págs. 595–598). hacer:10.18653/v1/W16-5903.
- Greenbacker, CF (2011). Hacia un marco para el resumen abstracto de documentos multimodales. En *Actas de HLT-SS '11 de la sesión de estudiantes de ACL 2011* (págs. 75–80).
- Greenbacker, CF, McCoy, KF, Carberry, S. y McDonald, DD (2011). Semántico modelado de documentos multimodales para resúmenes abstractivos. En *Actas de HLT-SS '11 de la sesión de estudiantes de ACL 2011* (págs. 75–80).
- Han, X., Lv, T., Hu, Z., Wang, X. y Wang, C. (2016). Resumen de texto usando modelo gráfico semántico basado en framenet. *Programación científica*.
- Harabagiu, S., moldavo, D., Morarescu, P., Lacatusu, F., Mihalcea, R., Rus, V. y Girju, R. (2001). Gistexter: Un sistema para resumir documentos de texto.
- Hennig, L., Umbrath, W. y Wetzker, R. (2008a). Un enfoque basado en la ontología para resumen de texto. En *Conferencia internacional IEEE/WIC/ACM de 2008 sobre inteligencia web y tecnología de agentes inteligentes* (págs. 540–544). hacer:10.1109/WIIAT.2008.175.
- Hennig, L., Umbrath, W. y Wetzker, R. (2008b). Un enfoque del texto basado en la ontología resumen. En *Conferencia internacional IEEE/WIC/ACM sobre inteligencia web y tecnología de agentes inteligentes* (págs. 291–294). hacer:10.1109/WIIAT.2008.175.
- Hípola, P., Senso, JA, Leiva-Mederos, A., & Domínguez-Velasco, S. (2014). Resumen de texto basado en ontologías. el caso de texminer. *Biblioteca de la Universidad de Columbia Británica: Emerald Insight*, 32(2), 229–248. hacer:10.1108/LHT-01-2014-0005.
- Issa, F., Damonte, M., Cohen, SB, Yan, X. y Chang, Y. (2018). Significado abstracto representación para la detección de paráfrasis. En *Actas de la conferencia de 2018 del capítulo norteamericano de la asociación de lingüística computacional: Tecnologías del lenguaje humano: 1* (págs. 442–452).
- Jobson, E. y Gutiérrez, A. (2018). Resumen de texto abstractivo usando atento rns de secuencia a secuencia. Informes de Stanford.
- Joshi, A., Fidalgo, E. y Alegre, E. (2018). Resumen de texto basado en aprendizaje profundo: Enfoques, bases de datos y medidas de evaluación.
- Joshi, M., Wang, H. y McClean, S. (2018). Gráfico semántico denso y su aplicación. en el resumen de un solo documento.
- Kasture1, NR, Yargal, N., Singh, NN, Kulkarni, N. y Mathur, V. (2014). un estudio de resumen abstracto usando representaciones semánticas e información a nivel de discurso. *Revista internacional de investigación en ciencia y tecnología emergentes*, 1 (6), 53–57.
- Katja, C. (2010). Compresión de oraciones múltiples: encontrar las rutas más cortas en gráficos de palabras. En *COLING '10 actas de la 23ª conferencia internacional sobre lingüística computacional* (págs. 322–330).
- Kan, A. (2014). Una revisión de los métodos de resumen abstracto. *Diario de Theo-Tecnología de la información crítica y aplicada*, 59(1), 64–72.
- Khan, A., Salim, N., Farman1, H., Khan, M., Jan, B., Ahmad, A., et al. (2018). abstracto-resumen de texto tivo basado en un enfoque gráfico semántico mejorado. *Revista internacional de programación paralela*, 1–25. hacer:10.1016/j.eswa.2016.01.007.
- Knight, K. y Marcu, D. (2000). Paso uno de resumen basado en estadísticas: Oración compresión. *AAAI-00*.
- Konstas, I., Iyer, S., Yatskar, M., Choi, Y. y Zettlemoyer, L. (2017). amr neuronal: Modelos de secuencia a secuencia para análisis y generación. En *Actas de la*

- 55ª reunión anual de la asociación de lingüística computacional (págs. 146–157). [hacer:10.18653/v1/P17-1014](#).
- Kurisinkel, LJ, Zhang, Y. y Varma, V. (2017). resumen abstracto multi-documento rización por extracción parcial de árboles, recombinación y linealización. En *8ª conferencia internacional conjunta sobre procesamiento del lenguaje natural* (págs. 812–821).
- Le, HT y Le, TM (2013). Una aproximación al resumen de texto abstractivo. En *Enterrancia conferencia nacional sobre computación blanda y reconocimiento de patrones (SoCPar)* (págs. 371–376). [hacer:10.1109/SOCPAR.2013.7054161](#).
- Lee, C.-S., Chen, Y.-J. y Jian, Z.-W. (2003). Extracción de eventos difusos basada en ontología agente de resúmenes de noticias electrónicas chinas. *Sistemas Expertos con Aplicaciones*, 431–447. [doi:10.1016/S0957-4174\(03\)00062-9](#).
- Li, W. (2015). Resumen abstracto de varios documentos con información semántica extracción de iones. En *Conferencia de 2015 sobre métodos empíricos en el procesamiento del lenguaje natural* (págs. 17–21).
- Liao, K., Lebanoff, L. y Liu, F. (2018). Representación de significado abstracto para multi-resumen del documento. [arXiv:1806.05655v1](#)
- Lin, J., Sun, X., Ma, S. y Su, Q. (2018). Codificación global para resumen abstracto. *En Actas de la 56.ª reunión anual de la asociación de lingüística computacional* (págs. 163–169).
- Liu, F., Flanagan, J., Thomson, S., Sadeh, N. y Smith, NA (2015). Hacia lo abstracto resumen utilizando representaciones semánticas. *HLT-NAACL*.
- Liy, P., Lamy, W., Bingz, L. y Wangy, Z. (2017). Decodificador generativo recurrente profundo para el resumen de texto abstractivo.
- Lloret, E., Teresa, M., Rom-Ferri y Palomara, M. (2013). Compendio: un resumen de texto sistema de marización para generar resúmenes de trabajos de investigación. *Ingeniería de Datos y Conocimiento*, 88, 164–175. [hacer:10.1016/j.datak.2013.08.005](#).
- dolfo Lozano-Tello, & Gómez-Pérez, A. (2004). Ontométrica: Un método para elegir la ontología apropiada. *Revista de gestión de bases de datos (JDM)*, 15(2), 18. Mohan, JM, Sunitha, C., Ganesh, A. y Jaya, A. (2016). Un estudio de ontología basado resumen abstracto. En *Cuarta conferencia internacional sobre tendencias recientes en informática e ingeniería* (págs. 32–37). [hacer:10.1016/j.procs.2016.05.122](#). Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T. y Sundheim, B. (2010). En-predicados retóricos para resúmenes cuasi-abstractivos. En *RIAO 2010* (págs. 138 y 139).
- Marsi, E. y Krahmer, E. (2005). Exploraciones en la fusión de oraciones. En *Procedimientos de la asociación de lingüística computacional (ACL)*.
- McDonald, RT (2006). Compresión discriminativa de oraciones con suave evolución sintáctica. *Identidad En 11ª conferencia del capítulo europeo de la asociación de lingüística computacional* (págs. 297–304).
- Mehdad, Y., Carenini, G., Tompa, FW y Ng, RT (2013). reunión abstractiva resumen con vinculación y fusión. En *Actas del 14.º taller europeo sobre generación de lenguaje natural*.
- Mehta, P. (2016). Del resumen extractivo al resumen abstracto: un viaje. En *Pro-actas de la 54ª reunión anual de la asociación de lingüística computacional* (págs. 100–106).
- Moawad, IF y Aref, M. (2012). Enfoque de reducción de gráficos semánticos para abstractivo resumen de texto. En *Séptima conferencia internacional de ingeniería y sistemas informáticos* (págs. 132–138). [hacer:10.1109/ICCES.2012.6408498](#).
- Modaresi, P. y Conrad, S. (2016). Simurg: un corpus multilingüe extensible para resumen abstracto de un solo documento. En *FUEGO'16* (págs. 24–27). [hacer:10.1145/3015157.3015161](#).
- Moratanch, N. y Chitrakala, S. (2016). Una encuesta sobre resúmenes de textos abstractivos. *En Conferencia internacional de 2016 sobre tecnologías de circuitos, potencia e informática* (págs. 762–765). [hacer:10.1109/ICCCT.2016.7530193](#).
- Munot, N. y Govilkar, SS (2015). Marco conceptual para resumen de texto abstracto. *marización Revista internacional sobre computación en lenguaje natural*, 4(1), 39–50. [hacer:10.5121/ijnlc.2015.4104](#).
- Nallapati, R., Zhou, B., dos Santos, C., Gulehre, C. y Lapata, M. (2016). Abstractivo resumen de texto usando rnns de secuencia a secuencia y más allá. *La conferencia SIGNLL sobre aprendizaje de lenguaje natural computacional*.
- Nápoles, C., Callison-Burch, C., Ganitkevitch, J. y Durme, BV (2011). paráfrasis-Compresión de oraciones tic con una métrica basada en caracteres: apretar sin eliminar. En *Actas del taller sobre generación de texto a texto monolingüe* (págs. 84–90).
- Nayeem, MT y Chali, Y. (2017). Fusión parafrástica para oraciones abstractas múltiples generación de compresión. En *CIKM '17 actas de la ACM 2017 sobre conferencia sobre gestión de la información y el conocimiento* (págs. 2223–2226). [hacer:10.1145/3132847.3133106](#).
- Nema, P., Khapra, MM, Laha, A. y Ravindran, B. (2018). *Atención impulsada por la diversidad modelo para resumen abstracto basado en consultas*. Biblioteca de la Universidad de Cornell. Nguyen, CQ y Phan, TT (2009). Un enfoque basado en la ontología para la expresión de frases clave tracción. En *Actas de los artículos breves de la conferencia ACL-IJCNLP 2009* (págs. 181–184).
- Nguyen, LM y Ho, TB (2004). Reducción de oraciones basada en ejemplos usando el modelo den markov. *Transacciones de ACM sobre procesamiento de información en idiomas asiáticos*, 1, 1–12.
- Niu, J., Chen, H., Zhao, Q., Sun, L. y Atiquzzaman, M. (2017). Absorción de varios documentos resumen tractivo usando gráfico de fragmentos y red neuronal recurrente. *IEEE icc 2017 SAC simposio pista de redes de datos grandes*.
- Oya, T. (2014). Resumen abstracto automático de conversaciones de reuniones. Oya, T., Mehdad, Y., Carenini, G. y Ng, R. (2014). Un resumen basado en plantillas resumen de la reunión: aprovechar las relaciones entre el resumen y el texto de origen. En *Actas de la octava conferencia internacional de generación de lenguaje natural* (págs. 45–53).
- Passonneau, RJ, Chen, E., Guo, W. y Perin, D. (2013). Puntuación piramidal automatizada de resúmenes usando semántica distribucional. En *Actas de la 51.ª reunión anual de la asociación de lingüística computacional* (págs. 143–147).
- Pawar, A. y Mago, V. (2018). Cálculo de la similitud entre palabras y sen-encias utilizando una base de datos léxica y estadísticas de corpus. *Transacciones IEEE sobre conocimiento e ingeniería de datos*, 1–14. [arXiv:1802.05667v2](#).
- Rachabathuni, PK (2017). Una encuesta sobre técnicas de resumen abstracto. En *Computación e informática inventiva (ICICI), conferencia internacional sobre* (págs. 762–765). [hacer:10.1109/ICICI.2017.8365239](#).
- Rananavare, LB y Reddy, PVS (2017). Una descripción general del resumen de texto. *En-Revista Internacional de Aplicaciones Informáticas*, 17(10), 1–17.
- Rossello, G., Basile, P., Semeraro, G., Ciano, MD y Grasso, G. (2016). Mejorando resúmenes neurales de textos abstractos con conocimientos previos. *URANIA-16*. Rush, AM, Chopra, S. y Weston, J. (2015). Un modelo de atención neuronal para abstractive resumen de oraciones.
- Sahooa, D., Bhoib, A. y Balabantaray, RC (2018). Enfoque híbrido de abstractivo resumen En *Conferencia internacional sobre inteligencia computacional y ciencia de datos (ICCIDS 2018)* (págs. 1228–1237). [hacer:10.1016/j.procs.2018.05.038](#). Singhal, S., Vats, A. y Karnick, PH (2018). Resumen abstracto neural. *Re-Puerto*. IIT Kanpur.
- Song, K., Zhao, L. y Liu, F. (2018). Mecanismos de copia infundidos de estructura para la abstracción. resumen tivo. En *Actas de la 27ª conferencia internacional sobre lingüística computacional* (págs. 1717–1729).
- Song, S., Huang, H. y Ruan, T. (2018). Resumen de texto abstractivo usando lstm-aprendizaje profundo basado en cnn. *Herramientas y aplicaciones multimedia*, 1(6), 53–57. [hacer:10.1007/s11042-018-5749-3](#).
- Tanaka, H., Kinoshita, A., Kobayakawa, T., Kumano, T. y Kato, N. (2009). Sintaxis-revisión de oraciones guiadas para resúmenes de noticias transmitidas. En *Taller de generación y síntesis de lenguaje* (págs. 39–47).
- Tran, T., Cimiano, P., Rudolph, S. y Studer, R. (2007). interpretación basada en ontologías ción de palabras clave para la búsqueda semántica. En *Conferencia internacional de web semántica* (págs. 523–563).
- UzZaman, N., Bigham, JP y Allen, JF (2011). Resumen multimodal de com-oraciones complejas. En *IUI'11* (págs. 595–598).
- Vadapalli, R., Kurisinkel, LJ, Gupta, M. y Varma, V. (2017). Ssas: Simulación semántica laridad para el resumen abstractivo. En *Actas de la octava conferencia internacional conjunta sobre procesamiento del lenguaje natural* (págs. 198–203).
- Viet, LD, Sinh, VT, Minh, NL y Satoh, K. (2017). Convamr: significado abstracto Análisis de representación para documentos legales. En *SCIDOC* (págs. 1–7).
- Wilca, GCV y Cabezedo, MAS (2017). Un estudio de resumen abstracto utilizando representaciones semánticas e información a nivel de discurso. En *Texto, voz y diálogo* (págs. 482–490). Saltador.
- Wang, L. y Ling, W. (2016). Generación abstracta basada en redes neuronales para opiniones y argumentos En *Actas de NAAACL-HLT 2016* (págs. 47–57).
- Xiang, C., Jiang, T., Chang, B. y Sui, Z. (2015). Ersom: una coincidencia de ontología estructural enfoque automático utilizando la representación de entidad aprendida automáticamente. En *Actas de la conferencia de 2015 sobre métodos empíricos en el procesamiento del lenguaje natural* (págs. 2419–2429).
- Yang, Q., Passonneau, RJ y de Melo, G. (2016). Pico: evaluación de la pirámide a través de au-extracción de conocimiento automatizada. En *Actas de la trigésima conferencia AAAI sobre inteligencia artificial* (págs. 2673–2679).
- Yao, J.-G., Wan, X. y Xiao, J. (2015). Resumen multilingüe compresivo basado en frases marización En *Actas de la conferencia de 2015 sobre métodos empíricos en el procesamiento del lenguaje natural* (págs. 118–127).
- Yao, J.-G., Wan, X. y Xiao, J. (2017). Avances recientes en el resumen de documentos. *Sistemas de Información del Conocimiento*. [hacer:10.1007/s10115-017-1042-4](#). Yousfi-Monod, M. y Prince, V. (2008). La compresión de oraciones como un paso en el resumen rización o una ruta alternativa en el acortamiento de texto. En *Coling 2008* (págs. 139–142).
- Yu, N., Zhang, J., Huang, M. y Zhu, X. (2018). Una red de operación para la abstracción compresión de oraciones tivas. En *Actas de la 27ª conferencia internacional sobre lingüística computacional* (págs. 1065–1076).
- Zajic, DM, Lin, J., Dorr, BJ y Schwartz, R. (2006). La compresión de oraciones como componente de un sistema de resumen de múltiples documentos. *Conferencia de comprensión de documentos de 2006*.
- Zhang, J., Zhou, Y. y Zong, C. (2016). Resumen abstracto entre idiomas a través del modelo de traducción fusión de estructura de argumento de predicado mejorada. *Transacciones IEEE/ACM sobre procesamiento de audio, voz y lenguaje*, 1–25. [hacer:10.1109/TASLP.2016.2586608](#).
- Zhang, X., Cheng, G. y Qu, Y. (2011). Resumen de ontología basado en oraciones rdf grafico. En *WWW 2007 / tema: Web semántica* (págs. 707–715).