Universidad Nacional Autónoma de México

Posgrado en ciencia e ingeniería de la computación

Representación de oraciones como series ponderadas con entropía para tareas de similitud semántica

Tesis

para optar por el grado de:
Doctor en ciencia e ingeniería de la computación

Presenta:
Ignacio Arroyo Fernández

Tutor principal:
Gerardo E. Sierra Martínez
Posgrado en ciencia e ingeniería de la computación
Tutor:

Juan-Manuel Torres-Moreno
Posgrado en ciencia e ingeniería de la computación

Ciudad universitaria, enero de 2019

Representación de oraciones como series ponderadas con entropía para tareas de similitud semántica

Tesis para obtener el grado de doctor en ciencia e ingeniería de la computación,
por la Universidad Nacional Autónoma de México

Ignacio Arroyo Fernández

Posgrado en ciencia e ingeniería de la computación

enero 2019

Agradecimientos

Me gustaría agradecer a las siguientes personas:

Al CONACyT por el apoyo brindado durante la realización de este trabajo (beca número 386128).

A mis papás y mis hermanos, a mi hijo Alan. A Janet y a mi nueva familia.

A las personas que conocí en este camino, que me abrieron sus puertas y que directa, indirecta o inversamente me enseñaron cosas tanto en el nivel personal como en el profesional.

Resumen

La representación vectorial de textos muy cortos en el nivel semántico de la lengua es un problema abierto en el área de Procesamiento de Lenguaje Natural (NLP, del acrónimo en inglés) y de la Inteligencia Artificial. Unos de los aspectos principales que hacen difícil este problema es que los métodos existentes requieren señales de supervisión. Estas generalmente se obtienen a partir de recursos anotados de diferentes tipos. Entre ellos se encuentra la anotación lingüística, las bases de conocimiento y las anotaciones de similitud semántica o de inferencia (p. ej. pares de preguntas y respuestas). Además, la mayoría de los métodos de representación del estado del arte están basados en máquinas de aprendizaje muy complejas que requieren cantidades enormes de datos de entrenamiento. Esto ultimo implica la necesidad adicional de alto poder de cómputo.

Por un lado, este tipo de recursos sólo están disponibles para dominios (o campos discursivos) que son demasiado generales (se conocen como de lengua general). Además, actualmente estos no están disponibles al menos para las lenguas más habladas en el mundo; la mayoría de los recursos de anotación solo se producen para el inglés y, en algunos casos, el español. Por otro lado, obtener una representación vectorial plausible del significado de una oración no es fácil debido a la complejidad estadística de las interacciones semánticas tanto entre palabras como entre los contextos que las rodean.

Con la finalidad de aportar al estado del arte en NLP, en esta tesis se presenta un método no supervisado de representación vectorial de oraciones. El método propuesto modela una oración como una serie ponderada de representaciones de palabras (*word embeddings*). Los coeficientes de la serie se ajustan de manera no supervisada mediante la entropía de Shannon, calculada a partir del evento conjunto de la ocurrencia de una palabra en una oración. La disminución de esta cantidad, significa que la palabra observada proporciona ganancia de información a cerca del corpus textual y de la oración observada. Se observó que el cálculo empírico de esta disminución de entropía se puede hacer usando el método conocido como TF–IDF (*"Term Frequency–Inverse Document Frequency"*).

El método propuesto, además ofrecer resultados competitivos con el estado del arte, ofrece otras ventajas significativas: modularidad, bajo costo computacional de entrenamiento, independencia de dominio, independencia de la lengua, independencia de conocimiento externo y de recursos de anotación lingüística. Asimismo, la implementación desarrollada permite generar representaciones de oraciones a bajo costo computacional y en tiempo real, ya que no requiere cargar en memoria ningún tipo de archivo (ya sea del modelo o de las representaciones).

Índice general

Índice de figuras

Índice de tablas

Introducción

Hoy en día, el crecimiento del volumen de información en los medios digitales da pie al análisis de grandes cantidades de datos de texto. Este tipo de análisis llama la atención de investigadores en ciencia de datos y en inteligencia artificial, así como de la industria de la Internet debido a que mucha información útil se puede extraer para diversas aplicaciones. Algunas de estas son muy lucrativas para las compañías que dominan la industria de la Internet (e.g. los datos textuales se pueden utilizar para el entrenamiento de algoritmos que asisten en el diseño de fármacos o de nuevos materiales para uso militar). Sin embargo, este tipo de aplicaciones requieren información específica y confiable cuya integridad, por ahora, solo un ser humano tiene la habilidad de verificar. Para ello se requieren años de entrenamiento y especialización que pueden no ser suficientes para analizar tantos datos disponibles.

En general, la información específica es difícil de detectar debido a que los usuarios humanos de la Internet son los principales contribuyentes de datos. Una motivación principal de estos usuarios para alimentar de datos a la red es el compartir conocimiento. Esto lo hacen escribiendo de una manera libre, asumiendo que el lector conoce la estructura de la lengua y el significado de una suficiente cantidad de sustantivos. A veces también se asume conocimiento del significado de expresiones idiosincrásicas de uso común (p. ej. *"moviendo el bigote"*). En temas especializados, quien escribe generalmente asume que el lector conoce el tema particular del que se trata. Así, aunque es basto el volumen de datos textuales disponibles, mucha información contextual es obviada durante la escritura (como el significado no literal de *"moviendo el bigote"*), dejando su decodificación y comprensión completamente a cargo del lector y del conocimiento previo que este posea sobre dicho tema al momento de la lectura. Desde el punto de vista del texto como única evidencia del proceso comunicativo, las asunciones que hace el hablante al escribir implican diversas ambigüedades (sintácticas o en cuanto al contexto de uso de palabras y frases específicas). Las computadoras se muestran aún poco capaces de lidiar con ellas. Los algoritmos para representación cuantita-

tiva de objetos de lenguaje natural pretenden ayudar con este tipo de problemas, los cuales no son poco complejos; son incluso difíciles de delimitar.

Una parte importante del conocimiento *latente* (o que potencialmente podría ser accedido o detectado) en el volumen de datos textuales se encuentra (y surge diariamente) en repositorios como foros, periódicos digitales y enciclopedias digitales. Estos repositorios son, en su mayoría, de acceso abierto. Tal apertura también impulsa el aumento en el volumen de los datos, lo que a su vez los hace cada vez más difíciles de consultar, ocultando el conocimiento más específico entre su propia bastedad. Como resultado, se tiene que los seres humanos somos capaces de producir más datos de los que podemos interpretar (Chen et al., 2005).

Un reto más específico debido al aumento en el volumen de datos textuales es el fenómeno de la duplicación de entradas a los repositorios mencionados (p. ej. la duplicación masiva de preguntas en los foros). Esto contribuye al incremento indeseable de información redundante. Otro reto surge cuando, por el contrario, se desea aprovechar la redundancia para evaluar la confianza de las noticias en los medios electrónicos de comunicación. Es claro que filtrar o verificar el conocimiento codificado en forma de datos textuales en estos repositorios es imposible de realizar manualmente para los seres humanos. Estos y otros problemas de procesamiento de datos textuales pueden ser abordados mediante algoritmos de representación de oraciones de acuerdo a su contenido, con lo que se puede determinar si estas son semánticamente similares (así, p. ej., se puede verificar si una pregunta ya fue previamente hecha en un foro cuya base de datos almacena millones de ellas). Este tipo de algoritmos constituyen el tema principal tratado en este trabajo de tesis, que a su vez se enmarca en el ámbito del Procesamiento de lenguaje natural (NLP, *Natural Language Processing*), una subdiciplina del la Inteligencia artificial.

Hablar del contenido de una oración puede tener distintas interpretaciones. En este trabajo de tesis se toma como tal al campo (o dominio) de la oración (o enunciación). Este consta de la terminología* propia de cada campo y en ellos se encuentran los términos de uso habitual de cada disciplina, profesión u oficio (Calsamiglia Blancafort and Tusón Valls, 1999). Por lo tanto, decimos que dos oraciones son de contenidos similares si contienen palabras propias o habituales de campos similares. En NLP, este tipo de tarea de comparación se conoce como medición de similitud semántica textual (STS, por sus siglas en inglés: *Semantic Textual Similarity*. "Similitud semántica" de aquí en adelante).

Actualmente, una opción para medir similitud semántica es mediante sistemas

*Aunque esta definición especifica que las unidades de estudio son *términos*, en este trabajo de tesis se tiene conciencia de tal especificidad, pero no se adopta. Por razones prácticas, el contenido de una oración, en este trabajo, se encuentra en las *palabras habituales* para cada campo o disciplina.

STS (Agirre et al., 2012, Hatzivassiloglou et al., 1999). Un sistema STS calcula una puntuación de similitud semántica (un valor real) entre pares de oraciones. La mayoría de los sistemas STS incorporan señales de supervisión entre las que se encuentran las bases de conocimiento, las enciclopedias electrónicas y los recursos de anotación lingüística, p. ej. tesauros y anotadores PoS (*Part of Speech*). Existen sistemas STS que son supervisados y requieren explícitamente etiquetas de similitud semántica anotadas por humanos (Mihalcea et al., 2006). Sin embargo, para lenguas diferentes al inglés (o incluso para textos de áreas de conocimiento especializadas), esos recursos no están disponibles o son muy escasos. Además, el alcance de tales sistemas se limita exclusivamente a medir la similitud semántica. Así, la representación vectorial de oraciones y, por lo tanto de su contenido semántico, quedan en segundo término o no se toman en cuenta.

En este trabajo se tiene como principal interés el de representar oraciones por su contenido semántico. Para observar la calidad de tales representaciones se hacen diversas pruebas de medición de similitud semántica entre ellas. La representación de oraciones (o cualesquiera fragmentos muy cortos de texto) consiste en embeberlas en un espacio vectorial tal que sus elementos (vectores) son estimadores de su significado y pueden representarse geométricamente sin supervisión, i.e. sin la necesidad de ejemplos de entrenamiento que representen un comportamiento dado por humanos y que las representaciones deban adoptar (Elman, 1991, Hinton et al., 1986). Puesto que tales vectores son representaciones del contenido de las oraciones que se pueden formar en una lengua, entonces deberían cumplir con el siguiente criterio de *consistencia*: los vectores deberían encontrarse en una vecindad pequeña del espacio vectorial que habitan (son geométricamente cercanos) cuando representan oraciones semánticamente similares. Por el contrario, cuando estos vectores representan oraciones semánticamente diferentes, entonces deberían estar alejados o no correlacionados en dicho espacio. La principal ventaja de este enfoque es que nos da la posibilidad de estudiar el comportamiento estadístico del significado en una lengua al nivel de oración para su procesamiento computacional. Otra ventaja es que las representaciones obtenidas pueden ser operadas en espacios euclidianos. Esto permite usar métodos tradicionales de comparación entre vectores, tales como métricas (distancias) y similitudes. Así, el rango de aplicaciones es poco restringido, pues la mayoría de estas asume espacios métricos para procesar oraciones por su contenido. Por lo tanto, las representaciones de estas son fáciles de aprovechar para su aplicación en múltiples tareas de NLP e Inteligencia artificial (además de STS) donde se requiera un vector representando el contenido de cualquier oración. Unos ejemplos de estas tareas son: agrupamiento semántico de oraciones, resumen automático de textos (Yu et al., 2017, Zhang

et al., 2012), clasificación automática de oraciones (Chen et al., 2017, Er et al., 2016, Kalchbrenner et al., 2014), detección de paráfrasis (Yin and Schütze, 2015), medición de similitud semántica (Boom et al., 2015, Yazdani and Popescu-Belis, 2013) o incluso análisis de sentimientos (Chen et al., 2017, Kalchbrenner et al., 2014, Onan et al., 2017). Estas tareas a su vez pueden ser parte, por ejemplo, de un sistema de NLP destinado a reducir la duplicidad de las preguntas formuladas por los usuarios de un foro. Este tipo de aplicación se incluye en la sección de experimentos de este trabajo (véase la sección 4.1).

En general, la aplicabilidad de cualquier método de representación depende principalmente de las necesidades de información que se desea satisfacer y de las facilidades que ofrezca el tipo de texto que se necesita embeber en un espacio vectorial (Arroyo-Fernández et al., 2018b, Salton and Buckley, 1988). El caso de la representación de oraciones se torna particularmente complejo, por un lado, a medida que la aplicación requiere información más detallada sobre el contenido de las mismas. Pero por otro lado, usar representaciones complejas para aplicaciones que requieren información superficial puede ocasionar inestabilidad y sobreajuste (*overfitting*) (Arroyo-Fernández et al., 2018a). Por ejemplo, en trabajos recientes se puede observar que representar oraciones para clasificar la polaridad de las emociones expresadas por los hablantes es menos complejo que representarlas para representar su contenido al grado de detectar si un hecho es negado, afirmado o dudoso. Los trabajos en NLP son generalmente de enfoques muy empíricos y no siempre se toma en cuenta que un indicador de la complejidad de cada uno de estos problemas es la dimensionalidad del espacio vectorial de embebido (Vapnik, 1998). A mayor dimensionalidad, mayor complejidad, y viceversa (Kalchbrenner et al., 2014, Yin et al., 2016). De hecho se puede observar en estos días una gran abundancia de trabajos que resaltan el alto desempeño de modelos basados en *Deep Learning* (DL) en tareas de clasificación de sentimientos. No así en tareas de medición de similitud semántica o representación del significado, donde métodos mucho menos complejos alcanzan resultados muy parecidos o mejores incorporando un poco de conocimiento lingüístico o *ingeniería de características* (Arroyo-Fernández et al., 2017) (algo difícil de inferir todavía para DL).

La complejidad de representación también puede variar con respecto al tamaño del texto representado. Por un lado, es un tanto directo obtener buenas representaciones de contenido para textos cuyo tamaño está en una escala relativamente grande (Le and Mikolov, 2014, Martin and Berry, 2007, Salton et al., 1983). Por ejemplo, los libros o cualquier colección de documentos que contengan cientos de miles de palabras o más, se representan bien mediante la importancia de las palabras que contienen. Un método de representación bien conocido y efectivo de este

tipo es la Bolsa de palabras (BoW) (Spärk Jones, 1972). Con este método es posible satisfacer necesidades de información relativamente superficiales que se limitan al tema del que tratan los documentos (Kintsch and Mangalath, 2011, Manning et al., 2009). Esto es útil, por ejemplo, en tareas de Recuperación de información y clasificación automática de documentos.

Por otro lado, las palabras están en la escala pequeña de los tamaños de texto. En esta escala, las necesidades de información que pueden surgir o ser satisfechas por un método de representación pueden ser muy generales. Las representaciones de palabras podrían ser componentes de prácticamente cualquier sistema de NLP. Estos métodos de representación se basan en un principio lingüístico conocido como la *hipótesis distribucional*. Esta hipótesis sugiere que las palabras con significados similares se usan en contextos similares (Firth, 1957, Harris, 1968). En el ámbito del NLP, este principio se modela como un estimador de la distribución estadística de los segmentos pequeños de texto de longitud fija (ventanas deslizantes de palabras) en los que aparece una palabra dentro de un texto grande. Este estimador codifica en varias dimensiones de un vector, tanto la distribución de los segmentos de contenido similar como la de los de contenido disímil. El vector que resulta de este cómputo estadístico para una palabra se conoce como *"Word Embedding"* (representación de palabra). Estos segmentos pequeños de texto en los que aparece una palabra se asumen como contextos. Estadísticamente, un estimador de estos es una media que representa a una familia de distribuciones de los contextos. Estos están contenidos en un conjunto grande y potencialmente complejo, lo que imposibilita a la media representar a todas las posibles distribuciones de contextos y a sus eventos atípicos. Esto constituye la principal limitación teórica de los métodos de representación vectorial, pero su impacto no es tan perceptible en aplicaciones de propósito general.

Para las aplicaciones de propósito general en NLP, las representaciones de palabras son suficientemente poderosas y simples. Una de estas aplicaciones es la generación de métodos de representación no supervisada de oraciones (Baroni and Lenci, 2010, Baroni et al., 2014, Bojanowski et al., 2016, Mikolov et al., 2013, Pennington et al., 2014). En esta tesis se evalúan distintos métodos disponibles de representación de palabras, como parte del método de representación de oraciones propuesto.

El método propuesto fue ajustado y evaluado mediante los conjuntos de datos comúnmente usados para STS en el concurso de evaluación semántica organizado por el *International Workshop on Semantic Evaluation* (SemEval), i.e. SICK (Jurgens et al., 2016), SemEval-2016 (Agirre et al., 2016) y SemEval-2017 (Cer et al., 2017). Los resultados obtenidos en dichas evaluaciones indican que dicho método fue

bastante competitivo con respecto a los métodos del estado del arte a pesar de ser muchísimo menos complejo. Ello no resulta sorprendente si se tiene en cuenta el fundamento teórico que le permite incorporar conocimiento previo en términos de los cambios de entropía dentro de las oraciones.

El método propuesto, además de ser no supervisado, ofrece ventajas significativas sobre los métodos del estado del arte: su implementación es modular, de bajo costo computacional de entrenamiento, es independiente del dominio o tema del texto representado, así como de la lengua. Además, no requiere recursos de conocimiento externo o recursos de anotación lingüística. Permite también realizar inferencia *"online"* de representaciones de oraciones a bajo costo computacional.

Planteamiento del problema

La representación de oraciones por su contenido semántico requiere información específica. Este tipo de información está codificado en diferentes regularidades estadísticas de la lengua, mismas que son difíciles de distinguir unas de otras debido a la pronunciada asimetría de las distribuciones de los objetos lingüísticos. Estos son completamente categóricos y el único número que puede establecer un orden entre ellos es la frecuencia con la que son observados y con la que se combinan para formar objetos más complejos. Cada objeto más complejo es observado con poca frecuencia, pero a su vez el conjunto de objetos complejos es bastante grande (e.g. sustantivos, términos, frases nominales, etc.). Así, cada uno de ellos es difícil de muestrear y caracterizar. Esto a su vez es un indicador de que cada objeto complejo transporta información específica (Chomsky and Schützenberger, 1963, Kuich, 1970). Pasa lo contrario con los objetos menos complejos, pues el conjunto de estos es muy pequeño (e.g. determinantes, preposiciones, verbos auxiliares, etc.). Al mismo tiempo cada uno de ellos es muy frecuente y puede ser observado prácticamente en cualquier oración. Se tiene con ello que dichos objetos no especifican información acerca del contenido de una oración, aunque muestrearlos sea trivial (debido a la frecuencia alta con que son observados).

Por lo anterior, si se desea que un método de representación de oraciones sea de propósito general (como en el caso de este trabajo), este debería mantener un equilibrio entre ser capaz de lidiar con la especificidad del contenido, mientras es también lo suficientemente general como para representar tantas oraciones como sea posible en una lengua. Es decir, que las representaciones de oraciones resultantes tengan un rendimiento alto y muy similar en tantos escenarios como sea posible. Por ejemplo, una necesidad de información con estas características puede ser "el saber qué se declara acerca de algo" (Collobert et al., 2011, Kintsch and

Mangalath, 2011, Pereira, 2000). Si esto mismo, o algo muy similar, es declarado por otra oración, entonces, en el espacio vectorial donde están embebidas, las representaciones de ambas oraciones deberían ser muy cercanas o correlacionadas también. En caso contrario, si un par de oraciones declaran cosas distintas sobre distintas cosas, entonces sus representaciones deberán estar muy alejadas también.

En el estado del arte sobre representación no supervisada de oraciones es posible encontrar métodos que pueden ser altamente dependientes de la aplicación y de su especificidad. En estos casos es difícil mantener su desempeño y comportamiento de manera uniforme/estable en varios escenarios (Pham et al., 2015). En otros casos, los métodos existentes alcanzan cierta estabilidad, pero son computacionalmente complejos y requieren grandes cantidades de datos de entrenamiento. Esto los hace poco transparentes como para estudiar sus propiedades en términos de la lengua (Kiros et al., 2015, Pagliardini et al., 2017). Estos problemas se vuelven críticos cuando sólo se tienen disponibles documentos de texto plano (datos no etiquetados) que además traten sobre temas muy especializados o que estén escritos en lenguas que no cuentan con recursos de anotación lingüística o de conocimiento.

Con la finalidad de abordar estos problemas y de aportar conocimiento nuevo sobre el problema de representación no supervisada de oraciones, en esta tesis se tiene la siguiente

Hipótesis. *Es posible representar oraciones o segmentos muy cortos de texto tomando ventaja del vínculo existente entre sus representaciones de palabras a partir de contextos (word embeddings) y la Información mutua entre cada palabra representada, el conjunto de oraciones que la contienen y el conjunto de todas las oraciones contenidas en un conjunto de datos textuales.*

Metodología

A efecto de confirmar la hipótesis planteada, se propone un modelo de representación de oraciones que consiste en una serie ponderada de representaciones de palabras (Arroyo-Fernández, 2013, Bojanowski et al., 2016, Levy and Goldberg, 2014, Mikolov et al., 2013, Pennington et al., 2014). Desde el punto de vista de cada palabra en una oración, los coeficientes de la serie correspondiente se ajustan de manera no supervisada mediante el cómputo de la Información mutua entre una palabra y un corpus donde esta es observada. Este corpus está contiene a su vez una estructura de conjuntos constituida por la propia oración representada y por el conjunto de oraciones que contiene a una palabra de interés. Esta estructura

se toma en cuenta al calcular la Información mutua en forma de descomposición para los conjuntos que la componen. El cálculo se lleva a cabo desde el punto de vista de la entropía de Shannon* (Kullback, 1997, Shannon, 1948, 1949, 1940). De esta manera, el comportamiento de los datos para representar oraciones otorga cierta libertad teórica para generalizar el enfoque y aplicaciones de este trabajo, tal como se muestra en el anexo D.

Para simplificar la implementación de este modelo, en este trabajo se usa la transformación TF–IDF (*Term Frequency–Inverse Document Frequency*) como aproximación (Aizawa, 2003, Spärk Jones, 1972). Con esto, es posible ponderar la contribución de información que cada palabra, mediante su representación, aporta a la representación de una oración. Es importante mencionar que este esquema de serie ponderada se puede ver como un sustituto de muy bajo costo computacional de los mecanismos de atención para redes neuronales recurrentes que actualmente se han retomado en variadas aplicaciones (Arroyo-Fernández and Meza Ruiz, 2017, Er et al., 2016, Fukushima, 1987, Hochreiter and Schmidhuber, 1997, Yin et al., 2016). Se ha nombrado al método que se propone es esta tesis como "Serie de información de palabras para embebido de oraciones" (*Word Information Series for Sentence Embedding*, WISSE) (Arroyo-Fernández et al., 2017).

Organización de esta tesis

Este trabajo de tesis está organizado de la siguiente manera. En el capítulo 1 se presenta el estado del arte sobre métodos no supervisados para representación de oraciones. Además se exponen algunos trabajos que sirvieron como exploración empírica sobre concentración de información en oraciones y sobre los elementos importantes de estas. En el capítulo 2 se explican los modelos que integran el método propuesto en esta tesis; empezando por métodos de word embedding. Después se expone una reinterpretación de TF–IDF en términos de Información mutua (en el sentido de la entropía de Sahnnon) para ponderar series de representaciones de palabras. Por último se define la similitud semántica y el criterio que se adopta para esta en este trabajo. Asimismo, se definen los métodos de evaluación que se usaron para determinar el rendimiento de las representaciones de oraciones que se obtuvieron.

En el capítulo 3 se explica a detalle el modelo de representación de oraciones propuesto en esta tesis, así como la interacción de los módulos que lo componen. En el capítulo 4 se presentan los experimentos propuestos y los resultados de

*Existen algunas otras definiciones de entropía como las de Rényi y Tsallis, todas derivadas de la ecuación de Boltzmann (Maszczyk and Duch, 2008, Rényi, 1961, Tsallis, 1988), pero estas no son estudiadas por ahora.

evaluación de nuestro método en tareas de similitud semántica. También se discute un ejemplo aplicativo donde se lleva a cabo agrupamiento jerárquico aglomerativo de segmentos cortos de texto que contienen contextos ricos en conocimiento. Por último, en el capítulo 5, se exponen las conclusiones de esta tesis. También se discuten las desventajas, de una serie de observaciones y de los trabajos futuros que se plantean a partir del modelo propuesto.

Capítulo 1

Representación de oraciones

El método de representación de oraciones que se desarrolla en esta tesis puede ser clasificado simultáneamente en dos categorías cuyo trabajo relacionado se reporta en este capítulo. La primera categoría se refiere en general a modelos y métodos de representación no supervisada de oraciones o fragmentos cortos de texto. Con estos métodos, el embebido de las oraciones en un espacio vectorial se realiza directamente desde el texto plano como conjunto de datos de entrada. Los principales métodos existentes de este tipo se basan ya sea solo en estadísticas de coocurrencia; o bien, en una combinación de estas estadísticas con redes neuronales artificiales.

La segunda categoría involucra a modelos del estado del arte que utilizan cualquier forma de sumatoria ponderada de representaciones de palabras (*word embeddings*). Estas se toman previamente construidas y sus coeficientes correspondientes se ajustan mediante aprendizaje, tanto supervisado como no supervisado. Nótese que en esta tesis se propone un método no supervisado, pero en este capítulo mencionamos también métodos supervisados basados en sumatorias ponderadas debido a que son muy parecidos al método que se propone. De hecho, es difícil hallar métodos no supervisados basados en sumatorias ponderadas.

1.1. Métodos no supervisados para representación de oraciones

En esta sección se describe brevemente una serie de métodos no supervisados cuyo objetivo es construir representaciones de oraciones para propósitos generales. Estos métodos no utilizan recursos externos o señales de supervisión, lo cual constituye un subconjunto de los principales rasgos del modelo de representación propuesto en este trabajo.

1.1.1. Term Frequency–Inverse Document Frequency (TF–IDF)

Un método popular de representación basado en estadística fue utilizado originalmente en aplicaciones de recuperación de información, donde se requiere representar documentos como vectores para posibilitar su recuperación a través de consultas. Estas a su vez se tratan como documentos que por lo tanto también estarían representadas como vectores. La representación más efectiva y simple fue por mucho tiempo la transformación TF–IDF (*Term Frequency–Inverse Document Frequency*) que en general normaliza las frecuencias tanto de las palabras como de los documentos de una colección. Esta normalización tiene por finalidad la de calcular qué tan importante es cada palabra en función de la frecuencia con la que es observada en un texto. La versión no normalizada de esta técnica, que se basa solo en frecuencias de palabras, se conoce como Bolsa de palabras (BoW, *Bag-of-Words*). Hoy en día la transformación TF–IDF sigue siendo muy usada y existen varias heurísticas alternativas a la idea original para aplicar transformaciones adicionales a las frecuencias de las palabras dentro de un documento, p. ej. presencia/ausencia de palabras (heurística binaria), logaritmo de la frecuencia (*sublinear*), etc. (Salton and Buckley, 1988, Salton et al., 1983).

La transformación TF–IDF da como resultado una *matriz término-documento* $X \in \mathbb{R}^{|d_j| \times |w_i|}$ tal que cada una de sus celdas x_{ij} representa un valor real positivo que es fución de: la frecuencia f_{ij} con la que una palabra w_i del vocabulario es observada en un documento d_j, del número total de documentos $|d_j|$ y del número de documentos $|d_{ji}|$ en donde w_i es observada dentro el corpus:

$$x_{ji} = f_{ij} \log_2 \frac{|d_j|}{|d_{ij}|} \tag{1.1}$$

Aquí, el término $\log_2 |d_j| - \log_2 |d_{ij}|$ se llama "frecuencia inversa de documentos" (IDF, *Inverse Document Frequency*) y se considera una constante de la colección de documentos. Dada su naturaleza monotónica (debida al logaritmo), la transformación (1.1) se traduce en un x_{ji} grande si w_i es relativamente rara o poco frecuente en una colección de documentos, i.e. se observa la tendencia de que $f_{ij} \to 1$ y $|d_{ij}| \to 1$. Por el contrario, x_{ji} resulta ser relativamente pequeña para palabras muy frecuentes (Spärk Jones, 1972). Como efecto general, esta transformación linealiza la distribución de ley de potencias (o ley de Zipf-Pareto) de las palabras en el corpus (Shawe-Taylor and Cristianini, 2004). Esta transformación, que por mucho tiempo fue tomada como una heurística, cuenta con algunas justificaciones teóricas. La primera, propuesta en (Shawe-Taylor and Cristianini, 2004), se basa en el efecto de linealización de la distribución de Zipf, de manera que TF–

IDF se puede replantear como una función kernel (*semantic kernel*). La segunda justificación se basa en Teoría de la información (Aizawa, 2003, Shannon, 1948). Este esquema permite ver a x_{ji} como una medida (o métrica) de la disminución en la entropía de un corpus $D \ni d_j$, dado que la palabra w_i es observada en un documento dado d_{ij}. Es precisamente esta formalidad la que se usa en este trabajo de tesis también como justificación teórica principal del modelo propuesto. Esto porque es más intuitiva en términos del discurso cognitivo y de lenguaje natural (Charniak, 1996, Hartley, 1928).

En los experimentos presentados en este trabajo (capítulo 4), se ha incluido TF–IDF para la representación de oraciones como un método de referencia base.

1.1.2. Latent Semantic Analysis

El Análisis de Semántica Latente (LSA, *Latent Semantic Analysis*) es una extensión de TF–IDF y por lo tanto de la bolsa de palabras (Landauer et al., 1998). Esta transformación toma como entrada una matriz de término-documento creada ya sea como BoW o bien como TF–IDF (Martin and Berry, 2007). Los vectores de esta matriz de entrada X representan documentos y se proyectan sobre los eigenvectores de esta, contenidos en una matriz U (esta matriz se puede ver de hecho con una matriz de embebido). De esta manera, el resultado de la transformación para cada documento de entrada (un vector disperso $d_i \in \mathbb{R}^n$) es su embebido $\hat{d_j} \in \mathbb{R}^k$, que es una versión densa, rotada, trasladada, ponderada y de menor dimensión ($k \ll n$). Las componentes de $\hat{d_j} \in U_k\Sigma_k$ son entonces los coeficientes de proyección de $d_j \in X$ sobre los k primeros eigenvectores $u_j \in U_k$ contenidos en U y su posterior rescalamiento debido a los valores singulares $\sigma_j \in \operatorname{diag}\Sigma$ (Golub and Reinsch, 1970). La descomposición en eigenvectores es obtenida usando el método Descomposición en valores singulares (SVD, *Singular Value Decomposition*):

$$X = U\Sigma V \approx U_k\Sigma_k V_k$$

El número de eigenvectores k de la matriz de embebido está asociado al número de temas que se cree están presentes en la colección de documentos. Esto es aplicable también al nivel de oración y se podría decir que, a este nivel, más que temas se tienen campos de enunciación (Calsamiglia and Tusón, 1999).

1.1.3. Métodos neuronales

El método llamado *Doc2Vec* (originalmente conocido como *Paragraph Vector*) utiliza una red neuronal para construir representaciones de propósito general para las oraciones o párrafos de una colección de documentos (Le and Mikolov,

2014). Doc2Vec usa representaciones de palabras previamente construidas* (Mikolov et al., 2013). La red neuronal recorre un corpus de texto plano usando ventanas deslizantes de palabras (contextos de longitud fija). Como las palabras de una oración son asociadas a sus representaciones, estas son utilizadas como evidencia Bayesiana $S_{B_i} = \{w_1, w_2, \dots\}$ para predecir cómo sería una representación virtual s_0 asociada a la oración. Visto de otra manera, el modelo de Doc2Vec aprende a asociar una representación s_0 a las representaciones de las palabras de la oración. Esta representación virtual no representa entonces a una palabra, sino más bien a la combinación de palabras en S_{B_i} dadas como evidencia:

$$P(s_0|S_{B_i}) = \frac{P(S_{B_i}|s_0)P(s_0)}{P(S_{B_i})}. \tag{1.2}$$

El entrenamiento de las representaciones de oraciones consiste en maximizar la probabilidad (1.2) para todas las i ventanas del corpus usando una red neuronal que adopta como parámetros libres a la representación s_0 y a las representaciones asociadas a cada palabra contenida en S_{B_i}. Nótese que, en tiempo de entrenamiento, una oración es en realidad una ventana de palabras; mientras que en tiempo de predicción una oración se pasa como entrada a la red neuronal y esta la toma como una ventana para la cual infiere un vector s_0. La mayoría de los métodos que se verán en esta sección están basados en esta idea, aunque con algunas variaciones.

Como una extensión de Doc2Vec, el modelo neuronal llamado *Skip-thought* (Kiros et al., 2015) produce representaciones de oraciones a partir de los estados ocultos de una Red neuronal recurrente (RNN, *Recurrent Neural Network* (Elman, 1991, Hochreiter and Schmidhuber, 1997)). En este marco, las dos oraciones a los lados de una oración central constituyen una ventana de contexto de dos elementos. La RNN recorre el corpus palabra por palabra, de manera que las representaciones de las palabras acaparadas por la ventana se proyectan en forma secuencial sobre la matriz de estados ocultos de la red. A cada ventana (vista como una oración) se le asigna una clase diferente (una clase de contextos). Una vez que la red fue entrenada de esta manera, es posible inferir la representación de una oración no antes vista por la red proyectando las representaciones de sus palabras en la matriz de estados ocultos. El ultimo resultado de estas proyecciones sucesivas se toma como la representación de la oración correspondiente (esto se puede ver como la predicción de un vector de contexto, asociado a las clases de contextos que la red aprendió).

Cabe señalar que la intuición que motivó este enfoque con RNN es que se piensa que con este tipo de modelos los patrones secuenciales inherentes al lenguaje

*Aunque la implementación de este algoritmo también construye representaciones de palabras, la fase de representación de oraciones asume tales representaciones como previamente construidas.

(el orden de las palabras) son tomados en cuenta; pues las RNNs son apropiadas para modelar secuencias (series de tiempo). Aunque este tipo de modelos ha sido muy popular recientemente, son máquinas de aprendizaje muy complejas, por lo que su costo computacional es alto y su desempeño ha sido bajo en tareas de similitud semántica (Arroyo-Fernández and Meza Ruiz, 2017).

La arquitectura llamada *FastSent*, propuesta por Hill et al. (2016), se basa en el modelo de Glove (Pennington et al., 2014) (ver sección 2.2.2) combinado con la arquitectura de Skip-Thoughts. Esta red combinada utiliza una matriz de parámetros que *codifica* en una sola representación la información de coocurrencia tanto de palabras como de oraciones (ventanas, en tiempo de entrenamiento). La arquitectura del modelo está diseñada para aprender representaciones de oraciones mediante una red neuronal SDAE (*Sequential Denoising Autoencoder*) y mediante la simulación de ejemplos negativos (*adversarios*). Este autoencoder genera un mapa entre las coocurrencias y los ejemplos negativos para aprender combinaciones de palabras que son probables e improbables como oraciones. Así, la representación de una oración no vista puede ser inferida como la capa intermedia (*the encoding*) de la red en oposición al modelo negativo aprendido y dado el conjunto de palabras de la oración.

El método de Wieting et al. (2016) propone una diferencia que en su momento ha sido significativa con respecto a la mayoría de los métodos de representación de palabras, aunque usando el mismo esquema de Doc2Vec. Este método, denominado *CHARAGRAM*, aprende representaciones de $n-$gramas de caracteres (Bojanowski et al., 2016, Mikolov et al., 2012). Para construir la representación de una palabra, estas representaciones de $n-$gramas son simplemente promediadas. Para construir representaciones de oraciones, se promedian las representaciones de palabras obtenidas. Con ello se obtiene el algoritmo de representación de oraciones llamado *CHARAGRAM-PHRASE*.

Otro método llamado *C-PHRASE* se basa en la estructura de dependencia de constituyentes sintácticos (Bentivogli et al., 2016). La idea subyacente a este modelo es muy similar a la propuesta por Levy and Goldberg (2014) para generar representaciones de palabras (sección 2.2.1). Es decir, la coocurrencia de las palabras está restringida por dependencias de función sintáctica, en lugar de por ventanas deslizantes de palabras.

El método denominado *Sent2Vec* es el muy similar a Doc2Vec (Pagliardini et al., 2017). Los autores extendieron Doc2Vec para considerar oraciones (ventanas de longitud dinámica), en lugar de ventanas de contexto de ancho fijo. Además, este modelo toma en cuenta $n-$gramas de palabras. Para ello debe aprender múltiples clases en su capa de salida. La arquitectura de Sent2Vec aprende una distribución

de tipos de palabras del vocabulario para cada oración con la que se entrena. Lo que equivale a la arquitectura *Skipgram* de Word2Vec (Mikolov et al., 2013) (sección 2.2.1).

Debido a la popularidad que actualmente las redes neuronales han alcanzado, en ellas se basa la mayoría de los métodos que se encuentran ahora en el estado del arte sobre representación de oraciones. A diferencia de tal enfoque, el método que se propone en este trabajo usa los modelos neuronales parcialmente (para obtener representaciones de palabras). La parte complementaria de este método está basada en medidas de entropía.

1.2. Métodos basados en sumatorias ponderadas

El primer acercamiento en cuanto a ponderación de representaciones de palabras para construir representaciones de oraciones fue introducido por Ji and Eisenstein (2013). Este acercamiento tuvo como aplicación la detección/clasificación de paráfrasis (similitud sinonímica de oraciones). Una reciente extensión de este trabajo se presentó en (Yin and Schütze, 2015). Los autores proponen aprendizaje supervisado de los coeficientes de cada representación de las palabras de una oración. Este método toma como coeficientes iniciales a los IDFs (las frecuencias inversas de los documentos donde ocurre cada palabra) de las palabras (véase ecuación (1.1)) compartidas por un par de oraciones. El método estima dos distribuciones complementarias de eventos. A saber, en primer lugar, una distribución que modela la ocurrencia de una palabra en dos oraciones que se sabe son paráfrasis. En segundo lugar, una distribución que modela la ocurrencia de una palabra en dos oraciones que se sabe que no son paráfrasis. La divergencia de Kullback-Leibler de ambas distribuciones es calculada para ponderar los IDFs iniciales. A partir de esta ponderación se obtienen nuevos coeficientes para las representaciones de las palabras compartidas entre las paráfrasis del conjunto de datos de entrenamiento. Una vez que se tienen las ponderaciones finales, se calcula la factorización negativa de las matrices de representaciones ponderadas para construir un nuevo vector de características que está asociado a cada par de oraciones. Como la factorización negativa es un problema de aproximación por mínimos cuadrados, la idea es que el vector de características sea a su vez un embebido de las diferencias (resaltadas por los IDFs) entre los pares de oraciones (Lee and Seung, 2001). En la parte final del algoritmo, cada vector es utilizado como entrada de un clasificador para identificar paráfrasis.

En (Zheng and Callan, 2015) los autores usan métodos de ponderación de representaciones para mejorar aplicaciones de Recuperación de información. Tanto

las representaciones usadas para representar documentos como las usadas para representar las consultas se ponderan mediante IDFs. El objetivo de ello es calcular la similitud promedio entre los embebidos de una consulta y los de documentos candidatos a ser recuperados. En (Kenter and de Rijke, 2015) se propone un enfoque similar pero supervisado. Los autores calculan una función de regresión a partir de un conjunto de datos de entrenamiento con el fin de predecir la ponderación de las representaciones de las palabras en un conjunto de prueba.

En (Boom et al., 2015) los autores proponen un enfoque supervisado para ajustar los pesos de las representaciones de las palabras que componen un fragmento de texto. El método representa textos de longitudes similares a la de un párrafo con 30 palabras o más. Los autores plantean un problema de clasificación binaria para determinar si cada par de un conjunto de pares está semánticamente relacionado o no relacionado (algo parecido a la detección de paráfrasis, pero menos estricto). Antes de calcular la sumatoria ponderada, este método reordena las representaciones de las palabras de cada segmento de texto de acuerdo con su valor de IDF asociado, aunque no es claro qué efecto o justificación tiene tal reordenamiento, ya que los IDFs no son considerados como parte del modelo. Los autores muestran también que las palabras sin importancia inducen sesgo al momento de calcular la similitud semántica entre pares de representaciones. Por lo tanto su modelo omite estas palabras de la representación de las oraciones. Una vez que las representaciones son ponderadas con los coeficientes aprendidos, se promedian para obtener la representación final del fragmento de texto. Los autores señalan que este promedio funciona bien para textos de aproximadamente 30 palabras o más. No obstante, proponen también modificaciones adicionales a su modelo para longitudes de texto variable.

Otro método relacionado es el propuesto recientemente en (Ferrero et al., 2017). Este método usa aprendizaje supervisado y un conjunto de entrenamiento STS para ajustar los coeficientes de las representaciones de las palabras de una oración. De manera similar a Ji and Eisenstein (2013), Yin and Schütze (2015), los IDFs son considerados como coeficientes iniciales que deben ser complementados. Tal complementación se realiza aprendiendo coeficientes para las partes de la oración (coeficientes PoS, *Part of Speech*) asociadas a cada palabra de una oración y un exponente global para el producto entre los IDFs y estos coeficientes PoS. Como resultado del aprendizaje se obtiene un coeficiente para cada palabra.

La mayoría de los métodos que se han presentado en esta sección son supervisados. Un método no supervisado lo propuso recientemente Arora et al. (2017). Sin embargo, con este método los coeficientes de la representación de una oración ajustan también con supervisión. Estos actúan como parámetros de distribucio-

nes multinomiales de coocurrencia. Este ajuste se lleva a cabo de acuerdo con la probabilidad de que una palabra aparezca junto con otras palabras dentro de una ventana de contexto. Nuevamente, usando la misma idea explicada para la ecuación (1.2) (Mikolov et al., 2013). Además se toman en cuenta aspectos parecidos a los tomado en cuenta para TF–IDF.

Lo anterior se logra mediante una combinación convexa entre la probabilidad de que una palabra ocurra en un contexto (una ventana deslizante) y la probabilidad de que esta ocurra en cualquier parte del corpus. Una posible ventaja sobre TF–IDF es la posibilidad de hacer entrenamiento por lotes (*mini-batches*) de los coeficientes de las representaciones. Se exponen a continuación los detalles principales de este método.

Los autores calculan cada coeficiente de la representación de una oración mediante el teorema de Bayes:

$$P(w_i|c_j) = \frac{P(c_j|w_i)P(w_i)}{P(c_j)}, \tag{1.3}$$

donde $P(w_i|c_j)$ es la probabilidad de que el contexto c_j sea *apropiado* para la palabra w_i. Si tal cosa se cumple, entonces $P(w_i|c_j)$ es grande y w_i recibirá una ponderación grande. Así, su contribución a la representación de la oración será grande. Por el contrario, si la probabilidad $P(w_i|c_j)$ es pequeña, entonces la representación de la palabra asociada recibirá una ponderación pequeña. Entonces, la contribución a la representación de w_i a la oración será pequeña. Puesto que la probabilidad (1.3) puede resultar grande para palabras muy frecuentes en todo el corpus, esta se pondera por mediante un *hiperparámetro* de compensación α (definido por el usuario) tal que el coeficiente final β_{ij} para cada representación en un oración j queda como:

$$\beta_{ij} = \alpha P(w_i|c_j) + (1 - \alpha)P(w_i), \tag{1.4}$$

donde $P(w_i|c_j)$ tiene función de densidad *softmax*, tal como se hace en Word2Vec (sección 2.2). Esto es:

$$p(w_i|c_j) = \frac{1}{Z} exp(x_{w_i}, x_{c_j});$$

con Z siendo una constante de normalización y las $x_{(\cdot)} \in \mathbb{R}^d$ siendo parámetros de la densidad. Estos últimos se consideran representaciones, tanto de c_j como de w_i. Y como en el caso de TF–IDF, $P(w_i)$ es una constante que se calcula a partir de todo el conjunto de datos $D \ni c_j$. La similitud con TF–IDF se observa en el hecho de que β_{ij} toma en cuenta a los contextos que comparten a w_i, a la oración como contexto observado y a su frecuencia en D. La versión supervisada de este método usa un conjunto de entrenamiento STS para ajustar los coeficientes de la

representación resultante, de manera muy similar a como se hace en (Boom et al., 2015, Ferrero et al., 2017).

A diferencia del enfoque de Arora et al. (2017), el que se presenta en esta tesis se basa en la intuición de que cada palabra de una oración contribuye con una cantidad de información determinada, la cual pondera la contribución de su representación a la representación de la oración que la contiene. Esto se puede interpretar también de manera complementaria. Cada palabra tiene la capacidad de disminuir la entropía de la oración en diferente medida. Por ello, para formar una oración a partir de representaciones de palabras, las más entrópicas deberán estar más penalizados que las más informativas.

1.3. Concentración de información en oraciones

Zellig S. Harris (1991) sugiere que la mayor concentración de información en las emisiones individuales de las lenguas naturales se da principalmente en los predicados de las oraciones (Meza-Ruiz and Riedel, 2009). Además, existen trabajos donde se han estudiado las variaciones de entropía en oraciones. Esta variación presenta patrones bastante regulares que tienen que ver, entre otras cosas, con el "esfuerzo" que requiere un hablante para comprender cada palabra de una oración a medida que la lee (Frank, 2013, Hale, 2003, 2006).

En este trabajo de tesis se estudia el problema de representación de oraciones teniendo en mente que es necesario ponderar los elementos separables de estas de tal forma que dicha ponderación sea proporcional a alguna medida de concentración de información. Desde el punto de vista de un estimador de tal medida, este debería proveer un mecanismo de ponderación estable, independientemente de la lengua y de los temas sobre los cuales tratan los textos analizados. Este razonamiento, a primera vista, parece poco complicado por estar muy relacionado con las partes de la oración. Esto porque mediante ellas es posible identificar unidades lingüísticas concretas tales como los predicados y por que las herramientas computacionales existentes para hacer esta identificación son bastante buenas. Aunque este resultado teórico y practico es sin duda cierto, existen todavía diversas limitaciones que dificultan su uso en escenarios de recursos lingüísticos escasos. A continuación se presentan algunos intentos de representación de oraciones que se exploraron durante el desarrollo de este trabajo y que fueron influenciados drásticamente por la hipótesis sobre la existencia de un método general para detectar concentración de información en oraciones.

Los experimentos exploratorios descritos en esta sección se consideran información empírica redundante y si relación directa con el modelo propuesto en este

trabajo, por lo que no se incluyen detalles sobre los modelos y métodos involucrados. Existen publicaciones que reportan con detalle todo lo relacionado con estas exploraciones empíricas y pueden consultarse en línea para fines de verificación y/o extensión (Arroyo-Fernández and Meza Ruiz, 2017, Arroyo-Fernández et al., 2016).

1.3.1. Dependencias sintácticas

De acuerdo al razonamiento antes mencionado sobre predicación e información, existen algunas formas de determinar qué elementos son semánticamente importantes en una oración (Wiemer-Hastings, 2005). Uno de los métodos más intuitivos es llevando a cabo mediante anotación automática de dependencias sintácticas. Este enfoque posibilita la detección de relaciones, .p. ej., del tipo sujeto-verbo -objeto (Manning et al., 2014).

Con este tipo de métodos, la información detallada (o concentrada) que se tiene es entonces, quién o qué realiza acción(es) sobre quién o qué. Este hecho se ha usado incluso para representar palabras (Levy and Goldberg, 2014). Sin embargo, se ha mostrado que este tipo de información incluso puede llegar a ser redundante para algunas tareas específicas de NLP en el nivel semántico (Arroyo-Fernández et al., 2018b), incluyendo la representación de oraciones por su contenido (sección 4.3). Nótese que en este caso se analizan relaciones asociadas específicamente a acciones. En los apartados siguientes se mencionan algunos experimentos exploratorios sobre representación de oraciones usando métodos basados en dependencias sintácticas.

1.3.2. Open Information Extraction

Existen métodos derivados de las dependencias sintácticas que hacen detección de *tripletas de información* (openIE, *open Information Extraction* (Angeli et al., 2015, Fader et al., 2011)). El atractivo inicial que se observó en estas herramientas consiste en que las tripletas constituyen explícitamente relaciones de acción. En términos de concentración de información, el uso de este tipo de conocimiento a cerca de una oración es bastante conveniente. Esto porque al identificar relaciones, es posible filtrar de manera muy puntual y directa toda aquella información que no sea específica para el mensaje codificado en una oración. Para observar con mayor detalle este razonamiento, se realizó un trabajo de investigación dedicado al uso de openIE para representación de oraciones. Este trabajo permitió comprobar las ventajas y limitaciones que presentan los métodos de openIE del estado del arte al ser utilizados en tareas de STS (Arroyo-Fernández and Meza Ruiz, 2017).

El problema fundamental en cuanto al uso de estos métodos es el *recall* bajo. Es decir, estos métodos son insensibles a la mayoría de los predicados que en realidad están presentes en un texto. Según un estudio reciente, los mejores métodos detectan exclusivamente relaciones verbales del tipo (NP, VP, NP), p. ej. "Juana$[NP]$ golpea a$[VP]$ Juán$[NP]$". Desafortunadamente, este tipo de relaciones comprende solo el 20 % de los predicados en un corpus de lengua general (Xu et al., 2013). Cabe señalar que en realidad este tipo de técnicas no fueron originalmente propuestas para el análisis de oraciones (análisis local), sino más bien para analizar grandes cantidades de documentos (análisis global), donde el corpus de estudio es la Web.

En teoría, los métodos basados en dependencias son independientes de los temas. Sin embargo, para textos muy especializados esto puede no ser cierto (p. ej. verbos muy propios de un campo de estudio pueden ser confundidos con sustantivos). La dependencia de la lengua es el principal inconveniente de este tipo de métodos debido a la necesidad de la construcción de un conjunto finito de reglas sintácticas (mismas que son válidas sólo para una lengua). No obstante, esto puede considerarse en segundo plano si la prioridad es detectar elementos importantes a partir de corpus grandes. Así, detectar el 20 % de predicados presentes en un corpus tiene sentido (se habla solo de aquellas relaciones que pueden ser pareadas por el conjunto de reglas sintácticas). Ello porque, generalmente, en todo fenómeno modelado por una ley de Zipf (e.g. las lenguas naturales) se cumple la regla $20 - 80$. Es decir, el 20 % de los patrones explica al restante 80 % (Mandelbrot, 1999, Newman, 2005).

Entonces, para un corpus grande como la Wikipedia, extraer el 20 % de las relaciones probablemente sea un buen muestreo (o estimación) de las relaciones totales o de la información un tanto específica que hay en él. Muy probablemente las relaciones importantes del corpus estarán expresadas al menos una vez en voz activa: sujeto-verbo-objeto, lo cual es fácil de detectar para un sistema openIE diseñado para el inglés. Esto, sin embargo, tiene pocas posibilidades de ser útil cuando se tienen intereses de estudio al nivel de oración. Por ello, y debido a la experiencia adquirida en (Arroyo-Fernández and Meza Ruiz, 2017), se concluye que el desarrollo actual de estos enfoques es por ahora inadecuado para inducir conocimiento previo en representaciones de oraciones.

1.3.3. Resumen automático

Otro acercamiento a la detección de elementos importantes de una oración consistió en realizar experimentos en resumen automático de documentos (ATS, *Automatic Text Summarization*) (Arroyo-Fernández et al., 2016). La motivación prin-

cipal para explorar este tipo de técnicas es la propia definición de la tarea que los métodos de resumen automático ejecutan. Esto es, ATS es el proceso de crear una versión reducida de un documento de texto, mediante el uso de software, de tal forma que esta nueva versión contenga los puntos principales del documento original. Así, surge la pregunta de si la idea de los puntos principales de un documento se puede extrapolar a los puntos principales de una oración de manera que al comparar los resúmenes de un par de oraciones se tenga un aproximado poco distorsionado de su similitud semántica.

Aunque durante el tiempo de desarrollo de este trabajo de tesis no fue posible contestar a esta pregunta, se hizo un primer acercamiento. Las dificultades principales de ello se debieron a la poca disponibilidad de métodos de ATS confiables y a que se verificó que su evaluación es poco fiable y muy subjetiva (cuestiones muy poco deseables que hacen, por ahora, prohibitivo el uso de técnicas de ATS en representación de oraciones).

En el acercamiento mencionado se planteó una tarea inicial en la cual se pretendía saber si un método de representación de oraciones del estado del arte embebe la información relevante que se desea saber sobre una oración. Para ello, se verificó si un modelo de Aprendizaje automático podía predecir la importancia de las oraciones embebidas. Los resultados mostraron que las representaciones utilizadas (Le and Mikolov, 2014) contienen información abstracta independiente del campo discursivo de la oración representada. Dicha información posibilita detectar patrones estadísticos en la lengua que caracterizan la importancia de la información presente en una oración.

El modelo de Aprendizaje automático (una SVM, *Support Vector Machine* (Cortes and Vapnik, 1995)) usó un esquema simple de transferencia (*knowledge transfer*). Este adquirió conocimiento a partir de una pequeña muestra de datos con una temática sin relevancia para el experimento (oraciones donde se usa la palabra francesa “puces” en contextos polisémicos). La muestra constaba de 30 oraciones manualmente calificadas según su importancia en un documento. Este entrenamiento fue transferido a una tarea de resumen automático de múltiples documentos (notas periodísticas) cuyas temáticas no estaban relacionadas con el documento de entrenamiento. Este experimento no fue relevante para la línea de investigación en ATS, pero sí fue concluyente en cuanto a la existencia de concentraciones de información en representaciones de oraciones obtenidas usando modelos neuronales. Por último, estos resultados sugerían también que aquello que genera dicha información puede ser identificado y reforzado para posteriormente aprovecharlo en representaciones de oraciones por su contenido semántico.

1.3.4. Redes neuronales recurrentes profundas con mecanismo de atención

Los métodos neuronales del estado del arte para representación no supervisada de oraciones ya se presentaron en la sección 1.1.3. A parte de esos métodos existen un tipo de redes neuronales especiales cuya característica es la detección de elementos importantes de un conjunto de secuencias o series de tiempo. Se trata de las Redes neuronales recurrentes (RNN, *Recurrent Neural Networks*; propuestas por Fukushima (1987)) con mecanismo de atención y las RNN con memorias de corto y largo plazo (LSTMs o *Long-Short Term Memories*, propuestas por primera vez por Hochreiter and Schmidhuber (1997)). A esta arquitectura combinada la llamamos aquí *attention- RNNs/LSTMs* y fue propuesta por Vinyals et al. (2015), quienes la aplicaron en tareas de anotación PoS. La motivación para explorar este tipo de métodos consiste en que resulta intuitivo ver a una oración como una secuencia de palabras tal que algunos de sus elementos requieren más atención que otros.

Se llevó a cabo un conjunto de experimentos exploratorios sobre aprendizaje de representaciones de oraciones usado la arquitectura attention-RNNs/LSTMs durante el desarrollo de este trabajo de tesis (Arroyo-Fernández and Meza Ruiz, 2017). Esto porque existían trabajos donde se observó un buen desempeño de las RNNs en la representación de oraciones (Kiros et al., 2015). Parecía intuitivo observar una mejora sustancial en la calidad de las representaciones usando mecanismos de atención en un escenario supervisado y donde se pudiera alimentar a la red neuronal con representaciones de palabras previamente entrenadas.

La arquitectura basada en attention-RNNs/LSTMs que se propuso consistía en dos de estas redes conectadas como gemelas (*twin networks*). Cada red gemela tenía dos capas LSTM (entrada), una capa oculta GRU (*Gated Recurrent Unit*, propuesta por Cho et al. (2014)) y una RNN plana (salida). Ambas salidas de las redes gemelas se combinaron con una red Maxout con capa de salida monolítica lineal (Goodfellow et al., 2013). Tanto las capas LSTM como las GRU estaban equipadas con mecanismo de atención. Las secuencias de entrada fueron representaciones FastText previamente entrenadas (Bojanowski et al., 2016). Esta red recurrente profunda fue entrenada de manera supervisada para estimar la regresión de las medidas de similitud anotadas por humanos, mismas que fueron provistas por los conjuntos de datos STS del SemEval (Cer et al., 2017). La arquitectura descrita fue la que mejor se desempeñó, pero se probaron otras combinaciones que no ayudaron a mejorar a pesar de su profundidad y complejidad computacional, p. ej.: solo RNN, solo LSTM, una LSTM + una RNN, una LSTM + dos RNN, etc.

Los resultados mostraron que el rendimiento de las representaciones obtenidas

mediante arquitecturas gemelas anttention-RNN/LSTM apenas fue comparable con el de las ya existentes (sin mecanismo atención) basadas también en RNNs, que ya era muy bajo (entre el 20 y el 30 %). Además, se confirmó el alto costo computacional reportado por Kiros et al. (2015). Entrenar este tipo de redes con 12000 pares de oraciones requiere al menos una semana de entrenamiento usando una GPU (el tipo de dispositivo de cómputo más rápido por ahora).

Indagaciones posteriores indicaron que cuando se aplican redes de atención a tareas de medición de similitud (en general, no necesariamente en texto), es necesario exponer su matriz de atención a las *interacciones entre los pares de objetos* que se comparan durante el entrenamiento. Esto constituye una configuración *siamesa* (Mueller and Thyagarajan, 2016, Neculoiu et al., 2016). Este enfoque requiere aprendizaje supervisado en un conjunto muy grande de datos con la finalidad de que el gradiente del error se propague por la matriz de atención al observar los contrastes entre los pares de muestras. Así, para el caso del modelado de similitud entre oraciones, la red neuronal "pone más atención" en determinados patrones contrastivos memorizados por la matriz de atención (Yin et al., 2016).

El método de representación de oraciones propuesto en este trabajo es no supervisado y se basa en series de representaciones de palabras ponderadas con entropía. Este mecanismo sustituye al mecanismo de atención descrito en esta sección, permitiendo así medir y controlar la contribución de información de cada palabra a la representación de la oración que la contiene. Es decir, las representaciones de las palabras más importantes en la oración (p. ej. sustantivos, verbos de acción y nominalizaciones) tienen un coeficiente mayor que permite ganar mayor cantidad de información (o atención, tal vez) a la representación resultante. Por el contrario, las palabras menos importantes (p. ej. artículos, preposiciones y verbos auxiliares) tienen un coeficiente mucho menor que bloquea su contribución mayoritariamente entrópica (en el sentido de que pueden ser observadas en cualquier contexto).

1.4. Motivación

Como se mostró en este capítulo, actualmente existen múltiples enfoques para la construcción de representaciones de oraciones en espacios vectoriales. Los más modernos pueden ser de mucha ayuda en aplicaciones de NLP de propósito general. No obstante, por un lado, la complejidad de las interacciones semánticas entre las palabras de una oración dificulta que estas representaciones se mantengan estables a través de múltiples escenarios. Esto es, la mayoría de los métodos funcionan relativamente bien al representar oraciones construidas de manera co-

rrecta o muy prototípica (sin ambigüedades lingüísticas) y para dominios muy generales. Sin embargo, la mayoría de los datos textuales generados en situaciones reales muestran pocos de estos beneficios, lo cual sigue siendo un cuello de botella en el área de NLP.

Por el otro lado, no existen (o son muy limitados) los recursos de anotación lingüística y de conocimiento para la mayor parte de dominios y lenguas. Aunque estos recursos son de mucha ayuda para insertar conocimiento a las representaciones, y así hacerlas más robustas a las expresiones atípicas en la lengua (Iacobacci et al., 2015), estos sólo están disponibles para dominios muy generales y lenguas mayoritarias (sobre todo para la lengua inglesa). Por lo tanto, los métodos no supervisados e independientes de recursos de conocimiento y anotación lingüística constituyen un nicho amplio de investigación.

Aunque en esta tesis se estudian los fundamentos de la representación no supervisada de oraciones, un avance significativo en este ámbito permitiría una gran expansión de aplicaciones de NLP limitadas actualmente por la falta de los recursos de anotación mencionados. Un gran número de estas aplicaciones se encuentra en dominios especializados y algunos de ellos son de impacto social. Por ejemplo, el dominio biomédico consta de literatura científica sobre investigaciones cada vez más abundantes de la biología aplicada a la medicina (el *bioNLP* se encarga de facilitar el aprovechamiento de esta literatura (Kim et al., 2009)). Otras aplicaciones numerosas que demandan métodos de NLP independientes de recursos de anotación son las relacionadas con compartir conocimiento en cualquier lengua natural, además del inglés.

Capítulo 2

Marco teórico

En este capítulo se muestran los diferentes métodos que se usaron o tomaron como referencia para implementar y evaluar el modelo propuesto en este trabajo. Primero se mencionan las consideraciones teóricas sobre lingüística computacional, luego sobre representación de palabras. Después se definen formalmente los sistemas y las tareas de STS. Con esto como referencia, se define un modelo general de representación de oraciones, así como la forma en que este se puede evaluar mediante tareas de STS. Por último, se expone una justificación teórica sobre TF–IDF en términos de Teoría de la información.

2.1. Composicionalidad en semántica distribucional

Mitchell and Lapata (2010) proponen modelos que se pueden considerar buenos candidatos para composición semántica en frases. Estos modelos se han probado empíricamente y ofrecen resultados prometedores. Entre ellos, la composición asimétrica tiene una forma particularmente interesante para este trabajo de tesis. Este modelo es una suma ponderada de representaciones de palabras. Su propósito es aproximar la composición de significado en frases cortas (p. ej., "variable aleatoria" y "va a querer"). En este marco, la asimetría se plantea como un rasgo lingüístico tal que la cabeza [h] de una frase es más importante que su modificador dependiente [d].

(1) variable[d] aleatoria[h].

(2) va a[d] querer[h].

El modelo de espacio vectorial para la composición asimétrica en frases como (1) y (2) está dado por:

$$p = \alpha x_{[\mathtt{d}]} + \beta x_{[\mathtt{h}]}, \tag{2.1}$$

donde se debe verificar que $\alpha < \beta$. Esta desigualdad refleja la diferencia entre la importancia de los constituyentes (las representaciones $x_{[d]}, x_{[\mathrm{h}]} \in \mathbb{R}^d$) de la frase resultante $p \in \mathbb{R}^d$. De Marcken (1999) estudió este fenómeno a detalle y al nivel de oración, en términos de un elegante modelo de fluctuaciones de entropía e Información mutua. Los coeficientes α, β son escalares que pueden estar dados por alguna función monótona, según Tian et al. (2017). En este trabajo de tesis se tiene la hipótesis de que una opción plausible para tal función es la entropía de Shannon (Aizawa, 2003, Charniak, 1996, Shannon, 1949). El modelo asimétrico además toma en cuenta tanto el orden de las palabras como las características estructurales que determinan la categoría sintáctica de las palabras que componen la frase resultante.

Dado que el modelo de composición (2.1) es transparente y natural, este ha alentado trabajo teórico reciente. Por ejemplo, Tian et al. (2017) introduce propiedades teóricas sobre el promedio y sobre la suma de vectores como operación de composición en semántica distribucional. A saber, dados un par de representaciones $x_{[\mathrm{d}]}, x_{[\mathrm{h}]}$ que no están correlacionados geométricamente, y si $\alpha = \beta = 1/2$, entonces se tiene que p se aproxima a cero. En otras palabras, la operación promedio hace que las representaciones de palabras que coocurren con baja frecuencia se cancelen mutuamente, por lo que $p \to 0$. Por el contrario, si se tienen representaciones de palabras que coocurren frecuentemente, estas generalmente están correlacionadas geométricamente (por ejemplo, "teléfono inteligente"). En este caso, la operación promedio favorece a la composición y p resulta como un nuevo vector en el mismo espacio donde habitan $x_{[d]}$ y $x_{[\mathrm{h}]}$. En este caso, p se puede ver como una representación que implícitamente está compartida por $x_{[d]}$ y $x_{[\mathrm{h}]}$.

Esta relación entre promedio, suma y correlación, teóricamente caracterizada por Tian et al. (2017), sugiere que en el subespacio vectorial de composiciones semánticas se tiene:

1. Tendencia a la dependencia lineal cuando el hablante expresa significados compuestos, pero lexicalizados o muy frecuentes.

2. Tendencia a la ortogonalidad entre representaciones cuando el hablante necesita componer significados menos obvios en el sentido de la frecuencia con la que se observan juntos sus constituyentes.

3. Anulación de la representación de la composición cuando sus vectores constituyentes no han sido observados conjuntamente (esto implica que estos vectores son prácticamente antiparalelos; o bien, que no están correlacionados).

A partir de nuestros experimentos, interpretamos que las observaciones antes mencionadas explican el bajo desempeño del promedio simple como método de combinación (o composición) de representaciones de palabras para representación de oraciones (sección 4.3.1). Por el contrario, esto mismo favorece la extrapolación del modelo de composición asimétrica hacia la representación de oraciones.

2.2. Representación de palabras (*word embedding*)

En esta sección describiremos los algoritmos que utiliza módulo de representación de palabras del método de representación de oraciones que se propone en este trabajo. Uno de los algoritmos de embebido más populares se denomina *Word2Vec* (W2V), propuesto por Mikolov et al. (2013), que es un modelo neuronal de lenguaje inspirado por ideas más generales sobre aprendizaje de representaciones propuestas originalmente por Bengio et al. (2003), Hinton et al. (1986). Recientemente se han propuesto métodos alternativos que se vuelven cada vez más populares; por ejemplo, *Glove* (Pennington et al., 2014) y *FastText* (Bojanowski et al., 2016).

Los métodos mencionados comparten características generales heredadas de una configuración de aprendizaje autodidacta (*"self-taught learning"*), la cual fue propuesta por Raina et al. (2007). En esta configuración, se entrena a un conjunto de estimadores cuyos parámetros son vectores aleatorios de valores reales. Estos estimadores modelan distribuciones multinomiales de la coocurrencia de cada palabra del vocabulario de un corpus con conjuntos de otras palabras, también del vocabulario (Rong, 2014). Esto es, dada una palabra $w(i)$, esta coocurre con un número determinado de otras palabras $w(i \pm 1), w(i \pm 2), \ldots$, encapsuladas dentro de una ventana deslizante o ventana de contexto $c_i = \{w(i \pm 1), \ldots, w(i \pm r)\}$. Al establecer esta regla para todas las palabras del corpus $i = 1, \ldots, |D|$, se producen estadísticas de coocurrencia que son utilizadas por el modelo de multinomiales para inferir el conjunto de palabras que con mayor probabilidad se encuentran en la vecindad c_i de la $i-$ésima palabra dada (véase la figura 2.1). En la mayoría de los casos de este tiempo, se usa una red neuronal cómo método de aprendizaje de estos estimadores, de manera que los parámetros de estos se vuelven parámetros de la red. Los métodos de embebido utilizados en esta tesis lo hacen así.

Una vez que el modelo está entrenado, sus parámetros son utilizados como representaciones de cada palabra del vocabulario. Estas representaciones pueden ser categorizadas en dos tipos, en general. Las primeras se conocen como *representaciones distribuidas* (Bojanowski et al., 2016, Mikolov et al., 2013) y resultan de la propagación del error de una red Perceptrón multicapa (en inglés *MultiLayer*

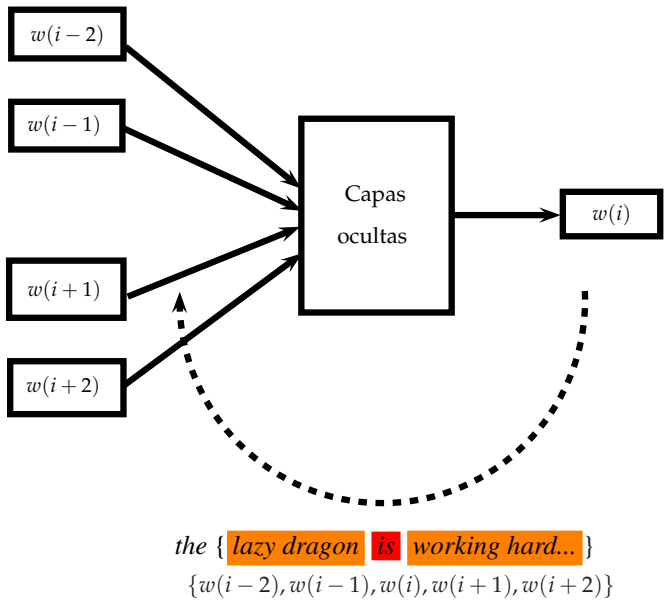

Fig. 2.1: Configuración *self-taught learning* usando una red neuronal

Perceptron [MLP], o también *Fully Connected Network*) hacia sus parámetros internos. Este error se toma a partir del procesamiento o aprendizaje independiente sobre las muestras de contextos del corpus no anotado de entrenamiento. Para Rumelhart et al. (1986), estos parámetros internos son *representaciones internas*. Otros métodos de representación ajustan sus parámetros a partir de una matriz dispersa previamente poblada por estadísticas de concurrencias. Estos modelos aprenden patrones de contraste entre palabras y contextos a partir de estas estadísticas (Dhillon et al., 2011, 2015, Pennington et al., 2014, Rastogi et al., 2015, Rothe and Schütze, 2016), lo que se asemeja en gran medida al enfoque de Raina et al. (2007).

2.2.1. Representaciones distribuidas

Word2Vec

W2V es un algoritmo de aprendizaje de representaciones de palabras basado en la configuración autodidacta. Es posible usar dos arquitecturas neuronales con W2V: Skip-gram y CBoW. En este trabajo, se ha utilizado Skip-gram para obtener representaciones.

Sea V el vocabulario de un corpus D, donde V contiene $|V|$ tipos y D tiene $|D|$ palabras (o *tokens*). La idea principal que subyace a Skip-gram es que clasifica

los tipos w_t a medida que estos toman lugar como $(w_i = w_t)$ a través del corpus $\{w_1, \ldots, w_i, \ldots, w_{|D|}\} = D$. Esta arquitectura es entrenada para predecir un conjunto de palabras $c_i = \{w_{i-r}, \ldots, w_{i+r}\} \setminus w_i$ (un contexto objetivo), dado que $w_t = w_i$ está en medio de ellas. Cada contexto objetivo contiene $2r = |c_i|$ palabras.

Antes del entrenamiento, un algoritmo de preprocesamiento transforma el texto plano de entrada sin etiquetar en un conjunto de entrenamiento de la forma $(w_1, c_1), \ldots, (w_{|D|}, c_{|D|}) \in D'$. En este nuevo conjunto D', w_i es una muestra de entrenamiento y c_i es su conjunto de etiquetas asociado. Una muestra se mapea hacia una distribución de múltiples etiquetas. Dado que las etiquetas c_i se obtienen a partir del mismo texto plano, el clasificador Skip-gram, un algoritmo de aprendizaje supervisado, se convierte en un algoritmo autodidacta (aunque no necesariamente no supervisado). La capa de salida de la red neuronal correspondiente puede escribirse en términos de sus capas ocultas de la siguiente manera:

$$P(c_i|w_i = w_t) = P(w_{i-r}, ...w_{i+r}|w_t) = \frac{\exp(x_t, \varphi_{c_i})}{\sum_{w \in V} \exp(x_t, \varphi_w)}. \tag{2.2}$$

Los vectores $x_t \in \mathbb{R}^d$ son los parámetros de la distribución de probabilidad $P(c_i|w_i)$ y se toman como representaciones de las palabras del vocabulario w_t. Estos parámetros pertenecen a la primera capa oculta de la red neuronal. La dimensión de las representaciones es d, que es igual al número de hiperplanos $\{(x_j, b_j) | \langle x_j, x_i \rangle + b_j = 0; x \in \mathbb{R}^{|V|}; j = 1, 2, \ldots, d\}$ descritos por la primera capa oculta de la red (x_i es un código disperso de w_i. Véase sección 2.2.3). Cada renglón de la matriz $\varphi_{c_i} \in \mathbb{R}^{d \times |c_i|}$ es el vector de coeficientes para cada nodo de la capa de salida correspondiente a cada tipo $w_{i-r}, \ldots, w_{i+r} \in c_i$. Cada renglón de la matriz $\varphi_w \in \mathbb{R}^{d \times |V|}$ corresponde a cada uno de los nodos de salida, es decir, a todos los tipos del vocabulario. Al reunir los elementos descritos se forma un MLP. A medida que la red es entrenada usando (2.2), esta analiza los datos de entrada, palabra por palabra. A partir de esto debe inferir las representaciones que parametrizan una distribución multinomial de los contextos en los que ocurren. De esta forma se busca que la distribución empírica de probabilidad de cada contexto c_i, dada una palabra w_i, converja a un valor esperado.

Representaciones basadas en dependencias (dep2Vec)

Otro de los algoritmos de embebido utilizados en este módulo es una extensión de W2V. Dicha extensión, que aquí llamamos *dep2Vec*, consiste en una modificación del concepto de ventana de contexto. Este algoritmo utiliza el análisis de dependencias como un sustituto de la ventana coocurrencias estándar c_i en (2.2). Dada una palabra w_i, el análisis de dependencias anota un enlace direccional entre

w_i y sus vecinos (Levy and Goldberg, 2014).

Se consideran vecinas, en una ventana de contexto, sólo a aquellas palabras $w_{i\pm r}$ que mantengan una conexión gramatical con w_i. Por ejemplo, la coocurrencia de la palabra "Juán" con la palabra "estornuda" es posible en este nuevo criterio de vecindad. No obstante, la coocurrencia de la palabra "mesa" con la palabra "estornuda" sería poco probable en un corpus de lengua general. Nótese que "Mesa" y "estornuda" es muy probable que coocurran dentro de una ventana de contexto usual (como en W2V), pero no tienen una conexión gramatical directa (o dependencia sintáctica). Por ejemplo, del tipo `sujeto-verbo`. En el aprendizaje de este tipo de representaciones distribuidas, se trata de inducir restricciones de contexto por dependencias. Ello, por ejemplo para (mesa, estornuda, Juán), limita las acciones que las personas y las mesas pueden realizar. Esto es, una persona estornuda, pero una mesa en realidad no hace tal cosa. Esta idea ya ha sido bastante explotada con anterioridad en los DSMs (Baroni and Lenci, 2010, Padó and Lapata, 2007). Recientemente se ha retomado con resultados relativamente buenos (Pham et al., 2015), aunque en representación de oraciones no hacen la diferencia con respecto a métodos menos complejos y libres de recursos de anotación lingüística (Arora et al., 2017).

Nótese que dep2Vec puede ayudar mucho en casos en los que se tenga una necesidad explícita de embeber preferencias de selección en representaciones de palabras (Baroni and Lenci, 2010). Sin embargo, sus autores de hecho muestran que la rigidez de esta restricción puede limitar la capacidad de las representaciones para codificar semántica de atributos. Por ejemplo, una versión estricta de esta restricción le impediría saber a las representaciones que *una persona*, Juán por ejemplo, frecuentemente estornuda *en la mesa*.

FastText

El algoritmo llamado *FastText* es otra modificación de W2V (Bojanowski et al., 2016). Este algoritmo es una fusión de las arquitecturas CBoW y Skip-gram. Es decir, dada una ventana de contexto $c_i = \{g_1, \ldots, g_{|c_i|}\}$ el algoritmo debe predecir una ventana objetivo $c_i' = \{g_1, \ldots, g_{|c_i|}, w_i\}$. Entonces se debe verificar que la probabilidad

$$P(c_i'|c_i) = P(g_1, \ldots, g_{|c_i|}, w_i|g_1, \ldots, g_{|c_i|})$$

converja para todo $w_i, c_i \in D$. Observe que ambos c_i y c_i' son un poco diferentes con respecto al c_i original (2.2).

FastText segmenta el texto de entrenamiento en n-gramas contiguos g_i de caracteres. De acuerdo con $|c_i|$, los n-gramas pueden pertenecer a otras palabras

que rodean a w_i. Además, las ventanas de entrenamiento c_i' contienen a la palabra objetivo w_i que está compuesta por n-gramas de caracteres g_i. Esto permite que el modelo genere diferentes representaciones para la preposición "con" y para el trigrama de caracteres *con*, contenido en la palabra "*con*vertir". El mismo mecanismo permite a FastText inferir representaciones fuera del vocabulario de entrenamiento, siempre y cuando los n-gramas de caracteres que la componen sí estén en el vocabulario de n-gramas de entrenamiento. Nótese entonces que este modelo aprende representaciones distribuidas de elementos estructurales de la lengua que son más básicos que las palabras. A este enfoque se le conoce como *subword information* (Mikolov et al., 2012).

2.2.2. Vectores globales para representación de palabras (Glove)

El algoritmo de Vectores globales para representación de palabras (Glove) es un poco diferente a la mayoría de sus contrapartes distribuidas. Este algoritmo retoma las ideas de los métodos pioneros como LSA. Su entrada de muestras debe estar en forma de una matriz dispersa de probabilidades de coocurrencias, calculadas a partir de una matriz de coocurrencias C (Pennington et al., 2014). Estas probabilidades son aprovechadas por Glove en tiempo de entrenamiento. Las entradas c_{ij} de C miden la probabilidad de que la palabra w_i coocurra con otras la palabras $w_1, w_2, \ldots, w_{|c_j|}$ que están al rededor de w_i en una ventana de contexto c_j. El modelo de entrenamiento calcula una función de regresión por mínimos cuadrados sobre las probabilidades mayores a cero. Los parámetros ocultos del modelo de entrenamiento (2.3) se toman como representaciones $x_i \in \mathbb{R}^d$.

$$\mathcal{J}(x_i, \varphi_j) = \sum_{w_i, w_j \in V} f(c_{ij})(\langle x_i, \varphi_j \rangle - \log c_{ij})^2 \tag{2.3}$$

Entonces, x_i corresponde a la palabra w_i, asimismo $\varphi_j \in \mathbb{R}^d$ es el embebido de la ventana de contexto que contiene a las palabras w_j. La función de penalización $f(c_{ij}) = \left(\frac{c_{ij}}{\text{máx}\{c_{ij}\}}\right)^{\alpha}$, un caso especial de la densidad de Pareto, contribuye para que el modelo de entrenamiento $\mathcal{J}(\cdot, \cdot)$ se adapte a cambios en la distribución monotónica de los datos. Este tipo de penalización se conoce como regularización ponderada (Sugiyama and Kawanabe, 2012).

Como muestra Pennington et al. (2014), $f(c_{ij})$ pondera la importancia de las coocurrencias. El objetivo convexo (2.3) optimiza el producto interno entre los parámetros libres x_i y φ_j inicializados aleatoriamente. Esto hace que el producto

entre ellos sea de valores cercanos o sea un estimador de $\log c_{ij}$, donde:

$$c_{ij} = P(w_j|w_i)P(w_i) : w_j \in c_j$$
$$\Rightarrow \log c_{ij} = \log \frac{P(w_j|w_i)}{P(w_i)} - \log P(w_i).$$

Esta última característica de Glove lo hace hasta ahora la mejor opción en la tarea de resolución de analogías de palabras (*word analogy*). Dicha tarea requiere generalización en la representación de palabras o frases. El alto rendimiento de este tipo de representaciones en esta tarea en particular no se mantiene al nivel de oración y para tareas de STS (véase la sección 4.3).

2.2.3. Dispersión en modelos de lenguaje neuronales

Los algoritmos de representación de palabras tienen una característica en común: la dispersión de las estadísticas de coocurrencia inducen ortogonalidad (Elman, 1991). En esta sección se presentan los principales casos durante el entrenamiento de un modelo de representación neuronal donde la ortogonalidad de las representaciones juega un papel importante.

En (Bengio et al., 2003, Mikolov et al., 2013) se utilizan vectores dispersos binarios para representar las palabras de un corpus como variables categóricas (*one-hot encoding*). Estos vectores construyen una base canónica para $X \subset \mathbb{R}^{|V|}$. Esta base a su vez codifica el vocabulario del corpus como un conjunto ortonormal

$$e = \{e_1, \ldots, e_{|V|}\} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \cdots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix}, \tag{2.4}$$

antes de iniciar el entrenamiento de las representaciones (Elman, 1991), que están en la capa oculta (o de proyección) del modelo y constituyen una transformación $T : X \to Y$, donde Y es cualquier espacio vectorial de salida (o de embebido).

Después del entrenamiento, la capa de proyección de la red neuronal del algoritmo de representación, ya ha codificado la ortogonalidad de las palabras de manera proporcional a la frecuencia con la que coocurren. Para palabras que coocurren frecuentemente, la ortogonalidad original en el espacio de coocurrencia $span(e) = \{0,1\}^{|V|} \subset X \subset \mathbb{R}^{|V|}$ es relajada en el espacio de embebido $Y \subset \mathbb{R}^d$. Esto hace que las representaciones correspondientes estén correlacionadas o sean linealmente dependientes en Y. Por el contrario, para palabras que coocurren muy poco (o que no lo hacen) la ortogonalidad inicial en $\mathbb{R}^{|V|}$ se transmite o se transfiere a las representaciones asociadas en Y.

Según Raina et al. (2007), este tipo de algoritmos se conoce también como de Codificación dispersa (*Sparse encoding*). Esto se debe a que el entrenamiento establece una relación entre un conjunto de vectores densos y un código disperso (o *Sparse Dictionary*) asociado a las componentes de la matriz e en la Eq. (2.4). La finalidad de esta relación es que los vectores densos representen elementos únicos que se puedan combinar para formar a e. Esta es también la finalidad de una base ortogonal en general. Pero para el caso de los algoritmos de codificación dispersa la ortogonalidad no es un requisito. En lugar de esto, el código disperso es un *marco redundante* o "sobrecompleto" (*overcomplete frame*), que es la generalización de las bases ortogonales (series de Fourier no armónicas) y cuya redundancia se debe a que el conjunto contiene elementos adicionales linealmente dependientes (Duffin and Schaeffer, 1952). Estos elementos redundantes tienen la finalidad de representar por separado patrones susceptibles de ser confundidos por efectos de ruido en los datos (Balan et al., 2006, Heil, 2007). Puesto que un marco redundante es un superconjunto de una base ortogonal (esto es, que está constituido por dos o más bases ortogonales), entonces sus elementos cumplen con la finalidad de combinarse para reconstruir un corpus mediante un vocabulario de representaciones densas (representaciones de palabras, en nuestro caso). Este hecho es de relevancia para el modelo de serie ponderada que se propone en este trabajo, ya que justifica que no se pida que los elementos de esta sean exclusivamente ortogonales. Así, se puede decir que el modelo propuesto en este trabajo, matemática y específicamente hablando, es una serie no armónica de Fourier cuyos elementos son coordenadas en una base redundante (que genera un *espacio vectorial de contenido semántico*) y sus coeficientes diferentes de cero (tantos como palabras tiene cada oración) la hacen una representación dispersa de las oraciones que se construyen con el método.

2.2.4. Valores extremos en el entrenamiento a partir de coocurrencias

En general se desea observar cómo los modelos de representación de palabras relajan la ortogonalidad de una matriz de coocurrencias, independientemente de si sus vectores son binarios o continuos o de si están basados en una red neuronal o no. Si para ello se considera el caso de Glove (Pennington et al., 2014), que no se basa en una red neuronal, y se consideran únicamente los valores extremos de sus parámetros que toman lugar durante el entrenamiento de este modelo, no es tan difícil exponer estas observaciones.

En Glove, se toma como entrada la probabilidad de que la palabra w_i coocurra con la ventana de contexto $C_j = \{w_1, \ldots, w_{i+r}\}$ (sección 2.2.2). Esto quiere decir que una matriz de probabilidades estimadas a partir de coocurrencias ya está po-

blada antes de entrenar del modelo. Entonces, dadas x_i y φ_j, las representaciones asociadas a w_i y a c_{ij} respectivamente, Glove optimiza la siguiente función objetivo:

$$\mathcal{J}(x_i, \varphi_j) = \sum_{i,j}^{V} f(c_{ij})(\langle x_i, \varphi_j \rangle - \log c_{ij})^2.$$

Si w_i no coocurre (o lo hace muy poco) con el contexto C_j, entonces $c_{ij} \to 1$ (c_{ij} se aproxima a 1) y por lo tanto se debe cumplir que $\log c_{ij} \approx 0$. Como el objetivo requiere que $(\langle x_i, \varphi_j \rangle - 0)^2 \to 0$, entonces el producto de punto $\langle x_i, \varphi_j \rangle \approx 0$. Esto implica que las representaciones x_i, φ_j tienden a ser ortogonales.

Una tendencia diferente puede observarse si w_i y C_j coocurren frecuentemente. El entrenamiento conduce a $\log c_{ij} > b$, de modo que $\langle x_i, \varphi_j \rangle \approx b$ para un $b > 0$ suficientemente grande. Por lo tanto, las representaciones x_i y φ_j tienden a ser linealmente dependientes en proporción al valor de b. Nótese que el logaritmo aplicado a c_{ij} penaliza las probabilidades de coocurrencia altas, mientras que favorece a las probabilidades bajas y maximiza a las probabilidades cercanas a la media. Esto sigue una relación entre $\langle x_i, \varphi_j \rangle$ y la entropía de Shannon.

2.2.5. Cuadratura en lugar de ortogonalidad

En el análisis de la sección 2.2.4, solo consideraron los valores extremos de coocurrencias y la media (inducida por el logaritmo). En general para todos los valores de coocurrencias, es posible observar una distribución de ángulos que se asemeja a la de un conjunto de armónicos esféricos, un conjunto de series logarítmicas (Jarosz et al., 2009, MacRobert, 1967), donde la ortogonalidad se vuelve relativa a la *cuadratura* entre los elementos del vocabulario (Sato et al., 1984, Sugiyama and Ogawa, 1999).

La cuadratura entre los elementos del vocabulario se representa en el plano complejo, mediante un diagrama de constelación similar al de la figura 2.2. Véase en la figura que en el mismo cuadrante (euclidiano) es posible codificar a más de un símbolo (e.g `110` y `111`). De manera análoga, se pueden observar símbolos que son ortogonales para los cuales se dice que *están en cuadratura* (e.g `001` y `100`). Asimismo, se pueden observar otros símbolos que mantienen correlación negativa entre sí (e.g `010` y `101`).

Tómese como ejemplo la red neuronal diseñada por Wen et al. (2016) para embeber un vocabulario de 10 símbolos en un espacio vectorial de 2 dimensiones, p. ej. el plano complejo. Si estos 10 símbolos fueran extraídos a partir de datos textuales, sabemos que algunos de ellos coocurren frecuentemente en un corpus. Se puede considerar entonces que estos se usan en contextos muy similares. Una

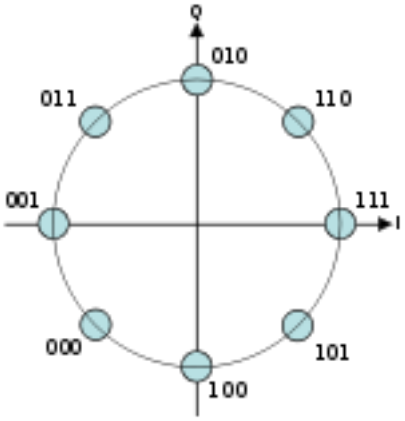

Fig. 2.2: Diagrama de constelación para un sistema de comunicación por modulación en cuadratura (QAM).

instancia de este patrón en procesamiento de imágenes se puede observar en la nube de puntos rodeando a c_5, en la figura 2.3.

Existen también símbolos que coocurren en contextos poco similares, como es el caso de los puntos que rodean tanto a c_4 como a c_5. Otros símbolos en cambio, deben ser codificados en cuadratura, como es el caso de (c_1, c_9), (c_1, c_8), (c_3, c_4), etc. Nótese que en estos casos particulares, los pares de contextos no son tan similares como para ser uno sólo; pero tampoco son tan disimilares como para mantener correlación negativa (como en los casos de, p. ej., (c_1, c_3), (c_8, c_9), etc.).

Dependiendo de la estabilidad del algoritmo de embebido, la cuadratura, la correlación y el antiparalelismo entre símbolos se mantienen en mayor o menor grado. Esto último se ilustra con las diferentes espacios de embebido que se generan al variar el parámetro de regularización λ del aprendizaje de la red neuronal del ejemplo de los 10 símbolos.

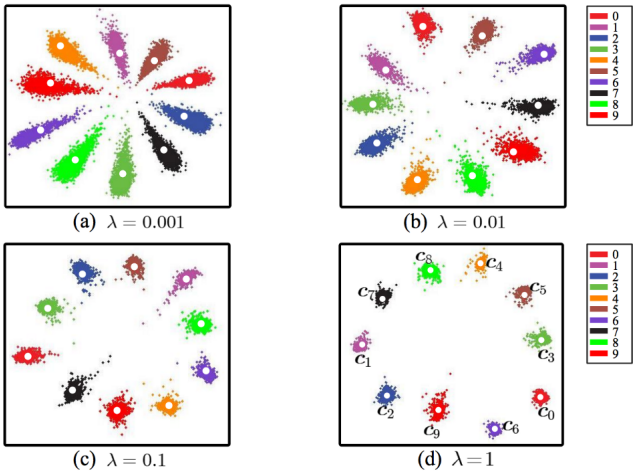

Fig. 2.3: Gráficas de representaciones aprendidas por una red neuronal en $\mathbb{R}^2$. Cada símbolo representado (dígitos del 0 al 9) en la capa oculta de la red está denotado como un círculo blanco (el prototipo del símbolo) en el centro de una pequeña nube de puntos. Cada nube de puntos es un conjunto de símbolos similares. La separación angular entre símbolos representaría la fase (o ángulo) de cada uno de ellos. El parámetro de regularización λ controla la estabilidad de la red neuronal. figura obtenida de Wen et al. (2016).

A partir de la figura 2.3 se puede ver que estos 10 símbolos son equiprobables, pues están rodeados cada uno por poblaciones de tamaños similares. Las clases están balanceadas, lo que cambia drásticamente en el caso de querer embeber los símbolos de una lengua natural. En este caso se tiene un problema bastante complejo de clases no balanceadas. Habría pocos símbolos con muchos elementos similares que los rodeen, pero muchos símbolos con pocos elementos similares a su alrededor. En los métodos actuales de representación de palabras, este comportamiento se intenta equilibrar mediante el submuestreo y el sobre muestreo de las clases (*subsampling*).

2.3. Teoría de la información en datos textuales

2.3.1. Point-Wise Mutual Information

Uno de los enfoques de representación de palabras más aceptados por mucho tiempo fue la Semántica distribucional (*Distributional Semantics*). En este marco, el significado de las palabras se representa en un Modelo de espacio vectorial (VSM, *Vector Space Model*) o Modelo de semántica distribucional (DSM, *Distributional Semantics Model* o también *Distributional Memory*). En años recientes, este último término fue el que más se adoptaba en investigación sobre representación

de palabras. Para construir un DSM a partir de un corpus, se requieren cantidades bastante grandes de datos, tantos como sea posible (Baroni and Lenci, 2010). Con estos se calculaban matrices de coocurrencia, donde los renglones corresponden a palabras de interés w_i que se desean representar y las columnas representaban otras palabras w_j del corpus con las cuales cada palabra de interés coocurre en una ventana deslizante c_j. Cada celda contiene frecuencias de coocurrencia (como en Glove. Véase sección 2.2.2).

La matriz de coocurrencia es transformada mediante *Point-Wise Mutual Information* (Jurafsky and Martin, 2014, Manning et al., 2009), de manera que ahora para cada celda se tiene:

$$PMI(w_i, w_j) = \log_2 \frac{P(w_i, w_j)}{P(w_i)P(w_j)}, \tag{2.5}$$

donde $P(w_i, w_j)$ es la probabilidad de que w_i coocurra con w_j dado que ambas están en c_j. Las probabilidades $P(w_i)$ y $P(w_j)$ se miden a partir de la frecuencia de ocurrencia de cada palabra en todo el corpus (lo que significa que también es probable observarlas independientemente). Puesto que la motivación original de la Información mutua es que sea una métrica de la información que comparten dos variables aleatorias (o procesos estocásticos), entonces se puede decir que (2.5) modela una comparación estadística entre dos procesos estocásticos que escogen palabras aleatoriamente del mismo corpus para mostrarlas en una ventana de contexto.

La matriz C obtenida a partir de (2.5) es muy dispersa y de miles o millones de dimensiones, por lo que se usa SVD (Singular Value Decomposition) para obtener una transformada $C = U\Sigma V$ con representaciones densas de cientos de dimensiones (*low-rank approximation*); esto es, una aproximación de menor dimensión k, i.e. $\widehat{C} = U_k\Sigma_k V_k$. Así, las dimensiones de esta matriz representan combinaciones lineales de las coocurrencias más informativas. En otras palabras, (2.5) es un estimador de la cantidad de información de cada coocurrencia. Por lo tanto, geométricamente las componentes de las representaciones de palabras de interés en $\widehat{C}$ son correlaciones con respecto a sus eigenvectores U_k, los cuales representan las direcciones en las que la cantidad de información (información mutua) de las coocurrencias registra mayor variabilidad. De esta manera, $\widehat{C}$ es una versión filtrada de C que mantiene solo las combinaciones más informativas de coocurrencias evitando así tanto las coocurrencias más obvias como las más entrópicas. Es esta información entonces la que se tiene en un DSM, en sus versiones más simples.

2.3.2. TF–IDF e Información mutua

En este trabajo se adopta una reinterpretación de la transformación TF–IDF, previamente definida en la sección 1.1.1, en el sentido de la Teoría de la información. Dado un vocabulario V contenido en un conjunto de segmentos de texto S (o corpus), se puede formar una estructura de subconjuntos de segmentos, donde las interacciones entre sus elementos se pueden modelar en términos del concepto de entropía de Shannon (1948). En esta estructura, los elementos directamente contenidos en el corpus son las oraciones $s \in S$. Después, cada oración contiene palabras $w \in V$.

Una propiedad importante de esta interpretación de TF–IDF en términos de información es que cumple las propiedades de la métrica de *Información mutua* de Shannon Kullback (1997) demostró que este tipo de métricas son convexas, lo que se deriva directamente del hecho de que la máxima incertidumbre en un proceso S se da sólo cuando produce muestras uniformemente distribuidas. Es decir, que todos los eventos s de S ocurren con igual probabilidad $P(S = s)$. Este comportamiento no se sostiene y tiene un desvío cuando se tiene conocimiento previo sobre S, digamos V:

$$H(S) \leq H(S|V) = H(S,V) - H(S),$$

donde

$$H(S) = -\sum_{s \in S} P(S = s) \log P(S = s)$$

es la entropía de S.

$$H(S|V) = -\sum_{s \in S;\ w \in V} P(S = s, V = w) \log \frac{P(S = s, V = w)}{P(V = w)}$$

es la entropía de S dado que V se conoce y

$$H(S,V) = -\sum_{s \in S;\ w \in V} P(S = s, V = w) \log P(S = s, V = w)$$

mide qué tanta incertidumbre se tiene sobre la ocurrencia simultánea los eventos de ambos procesos. Si los eventos de S ocurren simultáneamente y con cierta frecuencia (no con mucha frecuencia) con los de V, se tendrá una dependencia estadística informativa entre los procesos. La igualdad $H(S) = H(S|V)$ se da solo si S y V son procesos independientes, lo que lleva a $H(S|V)$ a un máximo acotado por $H(S)$ y por lo tanto no es posible saber algo sobre S a partir de V. Cuando se supone que V representa conocimiento previo sobre S, entonces $P(V = w) =$

$\sum_{s \in S} P(S = s, V = w)$ y por lo tanto, para observar el comportamiento de la entropía del corpus dado que se observa una palabra w en particular se puede escribir:

$$H(S|w) = -\sum_{s \in S} P(S = s, w) \log \frac{P(S = s, w)}{P(w)}.$$

Esto último representa el sesgo estadístico que en realidad ocurre entre subconjuntos de palabras en las lenguas naturales (y entre muchas otros elementos de la naturaleza) y que se da porque siempre se puede saber algo sobre el corpus (o disminuirá su entropía) al observar una palabra. Esto es, se cumple siempre que $H(S|w) < H(S)$. Usando estos resultados, Shannon (1948) y Chomsky and Schützenberger (1963) probaron que la distribución de los elementos de las lenguas es convergente y es estable. Esto se puede aprovechar para que los coeficientes de una serie ponderada por una métrica de información mutua converjan a un valor esperado:

$$I(S, V) = H(S) - H(S|V).$$

Es posible que por este motivo los resultados de STS, detallados en la sección 4.3, se mantienen relativamente estables en la mayoría de los escenarios de evaluación.

2.4. Información Mutua entre conjuntos de oraciones

Para calcular la Información mutua entre una palabra y el resto de subconjuntos de palabras que componen el corpus es necesario primero calcular la entropía de una oración. Retomando lo expuesto en la sección 2.3.2, sea una oración s_j. Suponemos que todas las $s_j \in S$ están uniformemente distribuidas, por lo que son igualmente probables: $P(s_j) = 1/N_S$, donde N_S es el número total de oraciones en el corpus S. Así, la entropía de cada oración s_j es la misma y se multiplica por N_S para calcular la entropía del corpus de oraciones:

$$H(S) = -\sum_{s_j \in S} P(s_j) \log P(s_j) \approx -N_S \left(\frac{1}{N_S} \log \frac{1}{N_S} \right) = -\log \frac{1}{N_S}, \tag{2.6}$$

donde $j = 1, 2, \ldots, N_S$.

Según Aizawa (2003), el siguiente paso es calcular la entropía del subconjunto de oraciones $S_i \subset S$ que comparten a w_i. Esto para medir cuánto es posible saber sobre el corpus dado que se escoge observar a w_i:

$$H(S|w_i) = -\sum_{s_j \in S} P(s_j|w_i) \log P(s_j|w_i) \approx -\log \frac{1}{N_{w_i}}, \tag{2.7}$$

donde N_{w_i} es la cardinalidad de S_i. El lado derecho de (2.7) asume que todos los elementos del conjunto $\zeta = \{S_i \subset S : i = 1, 2, \ldots, |V|\}$ están uniformemente distribuidos, por lo que son igualmente probables: $P(s_j|w_i) = 1/N_{w_i}$.

Finalmente, a partir de (2.6) y (2.7) se define la disminución en la entropía conjunta de S y S_i dado que w_i es observada. Esto es, el valor esperado de la información mutua que cada oración gana sobre ζ y S debido a que la palabra w_i es escogida en el proceso comunicativo (Kullback, 1997, Osteyee and Good, 1974):

$$\begin{aligned} I(V,S) &= \sum_{w_i \in V} P(w_i) \left[H(S) - H(S|w_i) \right] \\ &\approx \sum_{w_i \in V;\, s_j \in S} \frac{f_{ij}}{F} \left(\log \frac{1}{N_{w_i}} - \log \frac{1}{N_S} \right), \end{aligned} \tag{2.8}$$

donde $P(w_i) = \sum_{s_j \in S} f_{ij}/F$ es la probabilidad de observar a w_i en S, y F es la suma de las frecuencias de todas las palabras en S. Para una oración s_j dada, los elementos de la sumatoria (2.8) forman un vector de información.

$$x_j = \frac{1}{F} \left(f_{1j} \log \frac{N_S}{N_{w_1}}, \ldots, f_{|V|j} \log \frac{N_S}{N_{|V|}} \right),$$

que, como se mencionó en la sección 2.3, determina la importancia de cada $w_i \in s_j \in S$. Las componentes del vector $x_j \in \mathbb{R}^{|V|}$ serán utilizadas como coeficientes de una serie que representa a una oración en un espacio vectorial de contenido semántico de oraciones. Véase el capítulo 3.

2.5. Sistemas STS y métodos de representación de oraciones

Puesto que evaluamos nuestro método de representación de oraciones en tareas STS, en esta sección exploramos brevemente el contexto de estas tareas. Además, se muestra una comparación de rasgos distintivos entre métodos de representación de oraciones (como el propuesto en este trabajo) y de sistemas STS.

Existen conjuntos de datos para evaluar la predicción de similitud semántica entre pares de oraciones (i.e. conjuntos de datos STS). La mayoría de estos datos han surgido del concurso conocido como SemEval STS (Agirre et al., 2012). En este concurso, dado un par de oraciones, el objetivo es que un algoritmo venza a sus contrincantes al medir la similitud semántica entre ellas. Esta similitud es un número real que indica cuán semánticamente parecidas son las oraciones del par. Es decir, si estas dicen cosas muy parecidas o equivalentes; o bien si las oraciones del

par dicen cosas completamente no relacionadas. Cuanto mayor sea este número, mayor es la similitud semántica.

2.5.1. Sistemas STS

Con ayuda de los conjuntos de datos SemEval STS es posible evaluar el coeficiente de correlación entre la similitud calculada por algún algoritmo y la similitud anotada manualmente por humanos. Este coeficiente (el coeficiente de Pearson) es un número real $\rho(\cdot,\cdot) \in [-1,1]$. Por ejemplo, sea $\hat{y} = \{d(S_a^1, S_b^1), \ldots, d(S_a^\ell, S_b^\ell)\}$ el conjunto de similitudes calculadas por un algoritmo δ para ℓ pares de oraciones. También sea $y = \{y_1, \ldots, y_\ell\}$ el conjunto de similitudes manualmente anotadas por humanos. En este ejemplo, cuando $\rho(\hat{y}, y)$ se aproxima a 1.0 se tiene que el algoritmo de STS δ que calculó $\hat{y}$ se desempeña bien en la tarea de predecir las similitudes del conjunto de datos $(S_a^1, S_b^1, y_1), \ldots, (S_a^\ell, S_b^\ell, y_\ell)$. Se interpreta lo contrario en caso de que $\rho(\hat{y}, y) \to 0.0$. Una correlación negativa no necesariamente es mala en este tipo de competencias, pues, por ejemplo, $\rho(\hat{y}, y) \to -1.0$, significa que el algoritmo predice cosas opuestas y la interpretación de las predicciones simplemente se reinterpreta de manera acorde. Esto es bastante común cuando se usan distancias en lugar de similitudes para comparar oraciones.

Los sistemas STS del estado del arte alcanzan rendimientos relativamente altos ($\rho > 0.80$) (Cer et al., 2017). En este sentido, el problema de medir STS es relativamente maduro, al menos en casos usuales como cuando se tienen oraciones escritas en inglés general. Para otras lenguas y para temas específicos se dificulta mucho aún debido a lo escaso de los datos y de los recursos de anotación lingüística.

La mayoría de los sistemas STS actuales integran combinaciones de algoritmos que proporcionan medidas parciales de similitud a partir de diferentes atributos de las oraciones (Sultan et al., 2014). Estos pueden ser comparados tanto de formas supervisadas como no supervisadas y semisupervisadas. Por ejemplo, la Eq. (2.9)

$$d(S_a, S_b) = \alpha_1 d_1 + \cdots + \alpha_n d_n \tag{2.9}$$

constituye un método de mediciones parciales típicamente definido en la literatura sobre sistemas STS basados en *ingeniería de características* (Brychcın and Svoboda, 2016, Pilehvar and Navigli, 2015). En este modelo de mediciones parciales, los α_i son hiperparámetros. Estos también pueden ser ajustados de manera supervisada (usando un conjunto de datos STS). Cada parcialidad d_i es una medida de solapamiento entre los elementos del $i-$ésimo atributo (o categoría de atributos) alineado entre las oraciones del par S_a, S_b (p. ej. palabras alineadas por categoría

sintáctica o por su vecindad en un grafo de relaciones sintácticas o semánticas). Así, un modelo supervisado típico estima una función de regresión $d(\cdot,\cdot)$ a partir de los parámetros* $\alpha = (\alpha_1, \ldots, \alpha_n)$ de la Eq. (2.9). Para ello, se minimiza el funcional del riesgo de la Eq. (2.10):

$$L(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} |y_i - d(S_a^i, S_b^i)|^2 + \lambda \|\alpha\|, \tag{2.10}$$

donde $\lambda\|d\|$ es un término de regularización que limita la variabilidad del error (el término cuadrático de la ecuación).

Existen además sistemas semisupervisados cuya característica principal es el uso de recursos externos tales como bases de conocimiento, grafos semánticos, ontologías, tesauros y diccionarios digitales (Brychcın and Svoboda, 2016, Kenter and de Rijke, 2015, Pilehvar and Navigli, 2015). Otros métodos también incorporan información proporcionada por recursos de anotación lingüística como anotadores PoS, analizadores de dependencias sintácticas y analizadores semánticos. Incluso se pueden encontrar redes neuronales de las cuales se puede observar que, a pesar de su éxito en otras tareas, requieren de dicha información y de supervisión para alcanzar niveles de rendimiento competitivos (Brychcın and Svoboda, 2016, Rychalska et al., 2016). Un contraejemplo sin ingeniería de características sería la red convolucional siamesa de atención propuesta por Yin et al. (2016), aunque requiere supervisión y su costo computacional es alto.

2.5.2. Métodos de representación de oraciones evaluados en tareas STS

En este trabajo se toma ventaja de que para los métodos de representación de oraciones y para los sistemas STS es posible verificar su rendimiento usando las mismas tareas de evaluación (conjuntos de datos STS). En este trabajo se evalúa el desempeño del modelo de representación de oraciones propuesto mediante el uso de conjuntos de datos STS (i.e. SemEval-2016 (Agirre et al., 2016) y SICK (Bentivogli et al., 2016)). A pesar de estas similitudes, la interpretación de los resultados es un tanto diferente para evaluación de representaciones de oraciones.

Sea un método de representación de oraciones

$$\delta : X \to \mathcal{H} \tag{2.11}$$

que representa (o embebe) oraciones $S_{(\cdot)} \in X$, con X siendo un conjunto de sím-

*Cuando el modelo es no supervisado, los α son hiperparámetros, pues deben ser ajustados de manera manual. Sin embargo, cuando se tiene un planteamiento supervisado y que este aprenda los α directamente, estos se convierten en parámetros del modelo. En general, esta es la diferencia entre parámetro e hiperparámetros.

bolos y no necesariamente un espacio vectorial, en un espacio vectorial $\mathcal{H} \subset \mathbb{R}^d$ tal que $s_{(\cdot)} \in \mathcal{H}$. Entonces para un par de oraciones S_a, S_b, se tienen sus representaciones s_a, s_b, embebidas en $\mathcal{H}$ mediante la aplicación de δ. En el contexto de la evaluación, se desea saber qué tan bueno es el método $\delta(S_{(\cdot)}) \mapsto s_{(\cdot)}$. Se puede contestar a esta pregunta aplicando dicho método a tareas de STS. Con ello, la evaluación consiste en medir la similitud de las representaciones que embebe δ.

Esta similitud se mide de manera muy diferente a la empleada para sistemas STS (ecuación (2.9)). Para comparar representaciones de a cuerdo a la similitud de su contenido semántico se usa una medida de distancia o similitud definida en un espacio métrico (i.e., el espacio vectorial $\mathcal{H}$). En este caso, se tiene la la medida de similitud coseno

$$\hat{y}(s_a, s_b) = \cos(\theta) = \frac{\langle s_a, s_b \rangle}{\|s_a\| \cdot \|s_b\|}, \tag{2.12}$$

que puede ser usada como algoritmo de similitud semántica. Asimismo se pueden emplear distancias (o también llamadas métricas):

$$\hat{y}(s_a, s_b) = \left(\sum_{i=1}^{d} (s_j^{(a)} - s_j^{(b)})^p \right)^{\frac{1}{p}}, \tag{2.13}$$

donde $s_j^{(\cdot)}$ es la $j-$ésima componente de la representación $d-$dimensional $s_{(\cdot)}$ y $p < \infty$. En este trabajo se usan casos particulares de la métrica (2.13). Esto es, cuando $p = 1$ se usa la distancia Manhattan y cuando $p = 2$ se usa la distancia euclidiana.

Tanto la similitud (2.12) como las distancias (2.13) son funciones que se asumen algoritmos de medición de similitud y no se estudian a profundidad en este trabajo. Con esta premisa, el problema que demanda atención y esfuerzo de investigación es el método (2.11) para obtener las representaciones s_a, s_b que serán comparadas mediante dichos algoritmos. Así, el conjunto de similitudes $\hat{y} = \{\hat{y}(s_a^1, s_b^1), \ldots, \hat{y}(s_a^\ell, s_b^\ell)\} \equiv \{\hat{y}_1, \ldots, \hat{y}_\ell\}$ cuantifica la comparación por pares de representaciones de oraciones contenidos en un conjunto de longitud ℓ.

Dado el conjunto de mediciones de similitud $y = \{y_1, \ldots, y_\ell\}$ generado manualmente por humanos, para evaluar representaciones de oraciones se asume que estas similitudes son las reales. Por lo tanto, el objetivo de la evaluación es que el algoritmo de similitud, aplicado a los ℓ pares de oraciones, mida similitudes iguales a las provistas por los humanos en y.

Una forma de llevar a cabo el tipo de evaluación mencionada es mediante el

calculo del coeficiente de correlación ponderada (Eq. 2.14):

$$\rho(y, \hat{y}) = \frac{c_{y\hat{y}}}{\sqrt{c_y c_{\hat{y}}}}, \tag{2.14}$$

donde $c_{y\hat{y}}$ es la covarianza ponderada entre las similitudes reales y y las similitudes $\hat{y}$ medidas por el algoritmo $\hat{y}(s_a, s_b)$:

$$c_{y\hat{y}} = \frac{\sum_{i=0}^{\ell} w_i (y_i - \mu_y)(\hat{y}_i - \mu_{\hat{y}})}{\sum_{i=0}^{\ell} w_i}$$

donde

$$c_y = \frac{\sum_{i=0}^{\ell} w_i (y_i - \mu_y)^2}{\sum_{i=0}^{\ell} w_i}$$

y

$$c_{\hat{y}} = \frac{\sum_{i=0}^{\ell} w_i (\hat{y}_i - \mu_{\hat{y}})^2}{\sum_{i=0}^{\ell} w_i}$$

son las varianzas ponderadas de y e $\hat{y}$, respectivamente. La media ponderada de las similitudes reales es

$$\mu_y = \frac{\sum_{i=0}^{\ell} w_i y_i}{\sum_{i=0}^{\ell} w_i}$$

y la media ponderada de las mediciones del algoritmo es

$$\mu_{\hat{y}} = \frac{\sum_{i=0}^{\ell} w_i \hat{y}_i}{\sum_{i=0}^{\ell} w_i}.$$

El coeficiente $w_i \in [0, 1]$ se calcula en base al número de pares a comparar (entre más muestras, mayor es el coeficiente). Como un buen resultado de evaluación se espera que $\rho(\hat{y}, y) \to 1.0$. En este caso, se dice que el método de representación δ tiene un buen rendimiento en el conjunto de datos $(S_a^1, S_b^1, y_1), \ldots, (S_a^\ell, S_b^\ell, y_\ell)$ y con respecto a la similitud coseno como algoritmo de comparación. Adicionalmente, dicho resultado significa que tales representaciones $s_{(\cdot)}$ codifican una buena aproximación de los criterios humanos dados por y sobre el contenido de las oraciones.

2.5.3. STS contra métodos de representación

En general, los sistemas STS no necesitan representar oraciones en lo absoluto. De hecho, como ya se vio en parte en la sección 2.5.2, existe una notable distinción entre el objetivo de un sistema STS y el objetivo de un método de representación de oraciones evaluado en tareas STS. Esto es, la mayoría de sistemas STS están diseñados para tener buen desempeño en una tarea STS, mientras que se espera que

los métodos de representación de oraciones proporcionen buenas representaciones del contenido de las mismas. Esto permite prácticamente a cualquier sistema de cómputo localizar estos contenidos en un espacio vectorial entendible para él.

La principal característica de los sistemas STS es el uso de bases de conocimiento, p. ej. WordNet (Miller, 1995). Asimismo, el otro recurso preponderante lo constituyen los analizadores sintácticos, con los cuales se pueden hacer alineamientos de categorías de palabras. De estos sistemas, el más interesante, por su sencillez y efectividad, es el de Sultan et al. (2014).

Como ya se mostró en la sección 1.1, existen diversas dificultades en cuanto a costo computacional e interpretabilidad de los métodos del estado del arte. Por lo tanto, este trabajo está enfocado en proponer un método no supervisado de representación de oraciones que aporte avances en cuanto a estas dificultades y además, se espera que cumpla con características como la independencia de datos etiquetados y recursos de anotación lingüística, los cuales son externos al corpus de entrenamiento.

2.6. Resumen

Las series ponderadas que se proponen aprovechan las propiedades geométricas de la representaciones de palabras obtenidas mediante métodos neuronales. Teóricamente, es deseable tener un conjunto ortogonal como generador de un espacio vectorial de contenido semántico de oraciones (Arroyo-Fernández, 2013, Fourier, 1822). Sin embargo, se ha probado empíricamente que este enfoque es estricto y no siempre es el más efectivo para representar a los elementos de un conjunto de datos (Mairal et al., 2009, Raina et al., 2007). En este sentido, si las representaciones de palabras son un mapa que codifica al vocabulario, visto como un código disperso, entonces esto cumple con las funciones de generación del código a partir de los embebidos. Esto sugiere que no se requiere ortogonalidad estricta para generar un espacio vectorial de contenido semántico de oraciones usando series ponderadas. Por lo tanto, usamos este hecho para considerar a las representaciones de palabras como una base generadora de un conjunto de series ponderadas que se proponen en este trabajo como modelo de representación de oraciones.

Pennington et al. (2014) proporciona una prueba matemática de que la distribución de coocurrencias está acotada por un número armónico generalizado (es decir, una familia de distribuciones logarítmicas). Si se toman en consideración tal prueba y el análisis presentado en las secciones 2.1, 2.2.3 y 2.2.4, ello sugiere que los ángulos entre las representaciones de palabras también se distribuyen loga-

rítmicamente. Asimismo, se esperaría que se combinen (monotónicamente) para transmitir información en una serie ponderada. Esta distribución logarítmica de ángulos parece coherente con la entropía de Shannon y las estadísticas de coocurrencia. Esto también sugiere que un modelo de serie ponderada admite, por un lado, la contribución ponderada de representaciones que están moderadamente o poco correlacionadas, pero que juntos producen una composición informativa mediante la suma ponderada de sus significados.

Por otro lado, el modelo que se propone admite la contribución ponderada de representaciones de palabras muy correlacionadas, cuya combinación no es composicional. Más bien, representada por la resultante que tienen en común (frases lexicalizadas o muy frecuentes). Así, la serie ponderada de representaciones de palabras que se propone en esta tesis es un estimador que está integrado por dos subconjuntos no necesariamente disjuntos de ellas. El primero contiene unos elementos poco correlacionados que transportan información específica. El segundo subconjunto, más numeroso que el primero, contiene elementos que se correlacionan geométricamente, pero que transportan poca información. Se puede observar cómo esta idea se asemeja a la idea representada en el modelo asimétrico (2.1), pero en una jerarquía de *autosimilaridad** diferente (De Marcken, 1999, Mandelbrot, 1999).

*La autosimilaridad ocurre cuando un objeto es exactamente o aproximadamente similar a una o varias partes de sí mismo. En nuestro caso, las lenguas naturales son este tipo de objetos.

Capítulo 3

Serie ponderada por entropía de la información

En este capítulo se describe el método de representación de oraciones que se propone en este trabajo. Dicho método es una combinación de conceptos bien conocidos en la literatura de NLP (capítulo 2). No obstante de lo estudiados que ya están, la aportación principal de este trabajo es el estudio de estos conceptos desde puntos de vista teóricos y empíricos, tratando de no dejar de lado uno por otro. Esto con la finalidad de observar con mejor detalle la interacción entre los conceptos mencionados, que generalmente son estudiados de manera aislada; aunque se derivan de un mismo conjunto de datos. Además, los estudios existentes sobre los conceptos rescatados en este trabajo, desde el punto de Procesamiento de lenguaje natural, son de naturaleza predominantemente empírica.

El principal de estos conceptos es la información que un observador adquiere sobre un corpus dado que observa la jerarquía de subconjuntos contenidos en él (primero una palabra, la cual es observada en una oración, misma que está contenida en un conjunto de oraciones que comparte dicha palabra, y todo esto contenido en el corpus). Esto constituye una interpretación teórica de TF–IDF. El segundo concepto lo constituyen las representaciones asociadas a las palabras de una oración. Al contrario que en el caso de TF–IDF, la comunidad científica del mundo aún presta atención al desarrollo de modelos y métodos para este tipo de representaciones. Las evaluaciones de los métodos se hacen en tareas muy generales y bien conocidas al nivel de palabras como la similitud semántica (sinonimia) y la resolución de analogías semánticas y sintácticas (Arroyo-Fernández et al., 2018b). Sin embargo, hasta donde el autor de este trabajo ha explorado en la literatura, no existen evaluaciones de los métodos de representación de palabras en cuanto a su rendimiento al combinarse para representación de estructuras más

complejas que las frases (sección 2.1). Tampoco es posible hallar una discusión teórica, acompañada de pruebas empíricas, sobre tal rendimiento. Una estructura de interés inmediato es la oración, el cual es el objeto de interés de este trabajo.

Explotar el vínculo entre los conceptos mencionados, es decir, entre las palabras (como símbolos o variables categóricas), las propiedades estadísticas del lenguaje que las genera (Chomsky and Schützenberger, 1963, Kuich, 1970) y sus representaciones vectoriales, permite identificar vectores tanto informativos como entrópicos. Así se sabe cuáles de estos elementos son más importantes (informativos) y cuáles en realidad generan mayor incertidumbre (o son entrópicos) al representar la esencia del mensaje codificado en una oración. Este mecanismo automáticamente refuerza la contribución de representaciones informativas y disminuye la de aquellas que son entrópicas.

En cuanto a palabras informativas y entrópicas asumimos el siguiente punto de vista. El conjunto de palabras funcionales (p. ej. las preposiciones) no son palabras entrópicas por sí solas, pues la incertidumbre de que aparezcan en un contexto dado es muy baja. A su vez, se puede decir que son símbolos simples con poca combinatoria, lo que las hace muy probables de observar en cualquier corpus. Sin embargo, estimar su "significado" a partir de estadísticas de su uso está sujeto a mucha incertidumbre porque ocurren prácticamente en cualquier contexto. En otras palabras, de acuerdo con la hipótesis distribucional, estas palabras significan todos (o la mayoría de) los significados y en ese sentido son *ruido blanco**. Es desde este punto de vista que nos referimos a ellas como "entrópicas".

3.1. Modelo general

Aquello que se requiere saber sobre el mensaje codificado en una oración es información, la cual a su vez disminuye la entropía de este mensaje. Las muestras lingüísticas de diferentes longitudes construyen la estructura de la misma. De esta manera, y jerárquicamente, unas muestras proporcionan mayor información sobre el mensaje que otras, manteniendo así la entropía global del lenguaje en un límite contable y bien definido. Se dice que matemáticamente es un proceso convergente y estable (Chomsky and Miller, 1958, Chomsky and Schützenberger, 1963, Kuich, 1970, Shannon, 1948). El modelo general que subyace a WISSE† es una serie ponderada de representaciones de palabras que se beneficia de este princi-

*En física, la luz blanca es la presencia de todos los colores con la misma intensidad en una observación. Asimismo, la presencia de todas las frecuencias (de sonido) en una observación se conoce como *ruido blanco*.

†Acrónimo en inglés que se usó en la publicación original del método: *Word Information Series for Sentence Embedding*

pio matemático (Arroyo-Fernández, 2013, Arroyo-Fernández et al., 2017). De esta forma, las ponderaciones de la serie son calculadas a partir de la disminución en la entropía de un corpus a partir de cada palabra que se observa en la oración representada y contenida en el mismo (secciones 2.3.2 y 2.2).

Aquellas palabras de una oración que proveen más información sobre el contenido de la misma están asociadas a ponderaciones mayores. Por el contrario, las palabras que no permiten saber sobre el contenido de la oración están asociadas a ponderaciones menores. Sea A una oración vista como conjunto de objetos no ordenados, que en este caso son palabras $w_i \in A$. Entonces, en este trabajo se define a una representación para esta oración como

$$s_A(x, \varphi) = \sum_{w_i \in A} \varphi_{iA} x_{w_i}. \tag{3.1}$$

En el modelo general (3.1), los coeficientes de ponderación desconocidos $\varphi_{iA} \in \mathbb{R}$ regulan la contribución de cada representación $x_{w_i} \in \mathbb{R}^d$ de cada palabra w_i, dado que se observa en A. Así, la representación de A se denota $s_A(x, \varphi) \in \mathbb{R}^d$, donde $x = \{x_{w_1}, \ldots, x_{w_{|A|}}\}$ y $\varphi = (\varphi_{1A}, \ldots, \varphi_{|A|A})$.

Una intuición natural para conocer los coeficientes del modelo es optimizar el vector φ con respecto a un objetivo de aprendizaje. Un objetivo supervisado puede ser el error entre mediciones de similitud semántica anotadas por humanos y las predicciones hechas a partir de pares de representaciones de oraciones (véase sección 2.5.2). Otro objetivo puede ser autodidacta, usando la probabilidad de ocurrencia de cada palabra dado que las demás que contiene la oración ya fueron observadas juntas. Esto de hecho puede lograrse sin (o con poco) conocimiento previo acerca de los coeficientes (Arora et al., 2017, De Boom et al., 2016). Sin embargo, en este trabajo se adopta el enfoque de usar conocimiento previo que consiste en los patrones de entropía presentes en las oraciones que se forman en una lengua natural (Pereira, 2000, Schölkopf et al., 1997). Por un lado, las lenguas naturales contienen un conjunto de pocos elementos léxicos muy frecuentes (preposiciones, artículos, conjunciones, etc.). Los elementos de este conjunto tienen alta probabilidad de ser observados, así que decimos que son entrópicos (desde el punto de vista de los contextos en que ocurren). Por otro lado, las lenguas también contienen un conjunto muy numeroso de elementos, complementario del primero, que son relativamente poco frecuentes (frases verbales, entidades nombradas, sustantivos, etc.), por lo que la probabilidad de observarlos en una muestra lingüística es baja (Robertson, 2004). Decimos que la mayoría de elementos de este último conjunto numeroso son informativos, pues esta probabilidad de observarlos se acerca al valor esperado de elementos posibles. Estas propiedades estadísticas

de las lenguas naturales son muy conocidas, lo que ofrece la ventaja de que existe software para estimarlas y aprovecharlas en una aplicación.

3.2. Serie ponderada por entropía

La medida de información mutua considera la jerarquía de continencias entre conjuntos de palabras que se da cuando sucede el evento específico de que un hablante escoge una palabra w_i de su vocabulario V para construir la emisión de un mensaje u oración S_j. Véase a este último como un conjunto no ordenado, inicialmente. Mientras tal selección ocurre, al mismo tiempo hay otras palabras en V que, de manera natural, están adheridas a w_i. Uno puede imaginar esta adherencia como la viscosidad que se observa cuando se usa un rodillo de madera (o *cuchara mielera*) para tomar miel de un tarro. La palabra elegida se adhiere al rodillo. Esta a su vez se mantiene en cohesión con otras palabras que probablemente (o *viscosamente*) la acompañaran en el contexto (u oración) que se emite. Estas serán elegidas después con alta probabilidad. Otras palabras a su vez siguen adheridas a la superficie de miel que permanece el tarro, al rededor de la que fue seleccionada, pero están mucho más alejadas de ella. Véase la figura 3.1. La gran mayoría de las palabras restantes no fueron atraídas por el rodillo (aquellas que representan la miel en el fondo del tarro, p. ej.). Este último conjunto total se puede ver como un conjunto S de oraciones que probablemente el hablante escogerá decir en algún momento. En general es imposible estimar cuál de todas ellas es más probable de ser emitida sin antes observar una ocurrencia del vocabulario. Esta metáfora es de ayuda para imaginar los elementos de un módulo de entropía cuyos principios teóricos se definieron en la sección 2.4:

$$I(V,S) \approx \sum_{w_i \in V;\ S_j \in S} \frac{f_{ij}}{F} \left(\log \frac{1}{N_{w_i}} - \log \frac{1}{N_S} \right). \tag{3.2}$$

Como se dijo en la sección 2.3.2, los elementos de la sumatoria (3.2) se pueden estimar mediante la transformación TF–IDF. Esta sumatoria, expandida para $w_i \in V$ sobre una S_j fija, y el modelo general (3.1) son series que coinciden en su número de elementos para $f_{ij} > 0$, i.e. los términos de la sumatoria para los cuales $w_i \notin S_j$ simplemente desaparecen puesto que $0F^{-1}(\log 1/N_{w_i} - \log 1/N_S) = 0$. A partir de los términos que sobreviven, se define el vector TF:

$$\varphi_{S_j} = \left(\log \frac{1}{N_{w_1}}, \ldots, \log \frac{1}{N_{w_{|S_j|}}} \right) - \log \frac{1}{N_S}, \tag{3.3}$$

donde N_{w_i} es el número de oraciones que comparten a w_i y N_S es el número total

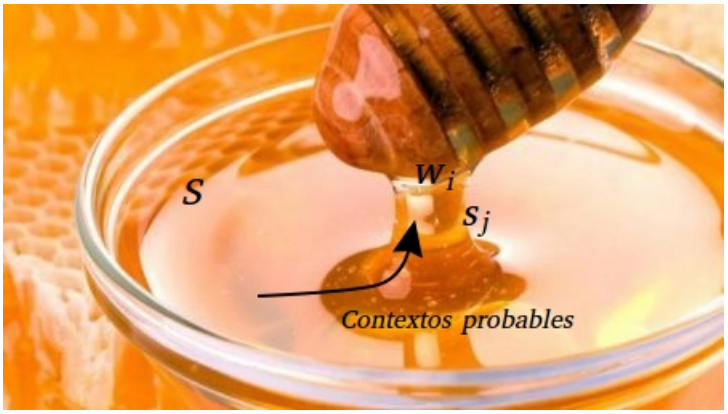

Fig. 3.1: Metáfora de la cuchara mielera y el tarro. Selección de una palabra, que a su vez sugiere los contextos probables para construir una emisión.

de oraciones en S. Nótese que cada componente de $\varphi_{S_j} \in \mathbb{R}^{|S_j|}$ proporciona la medida de decremento en entropía del corpus, dado que la palabra w_i es observada. De manera subyacente a la estructura de los objetos escogidos durante el proceso comunicativo, este decremento en entropía está ponderado por la probabilidad de observar a la palabra w_i en la oración S_j. Como lo indica (3.2), esta ponderación está determinada por la probabilidad

$$P(w_i) = \sum_{S_j \in S} P(w_i|S_j). \tag{3.4}$$

Cada término de (3.4) se calcula mediante

$$P(w_i|S_j) = \frac{f_{ij}}{F},$$

donde f_{ij} es el número de veces que w_i ocurre en la oración S_j y F es el número de tokens en el corpus.

Al estimar $P(w_i)$ a partir de este criterio, se considera la estructura (o jerarquía) subyacente a los subconjuntos de palabras que se han identificado en el mismo:

$$w_i \in S_j \subseteq \zeta_i \subseteq S,$$

donde se tiene la $j-$ésima oración S_j, misma que está contenida en el conjunto de oraciones ζ_i que comparten a w_i. Y estos conjuntos están contenidos a su vez en el conjunto S de todas las oraciones del corpus. Ignorar esta ponderación resulta en una omisión conceptual debido a que el teorema de Bayes define dicha ponderación para probabilidades condicionales. Por ejemplo, en la figura 3.2 se observa que para realizar cualquier cálculo sobre la probabilidad de un evento conjunto

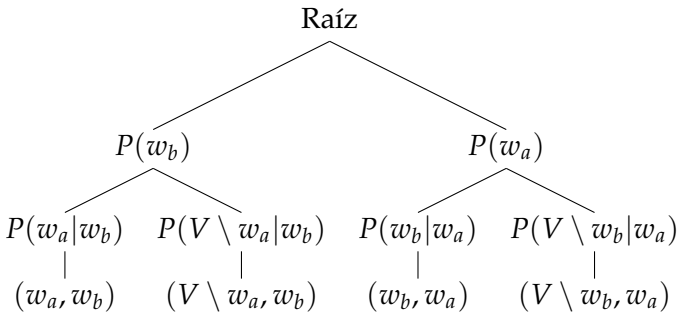

Fig. 3.2: Esquema del teorema de Bayes: estructura de probabilidades condicionales ponderadas.

(w_a, w_b) es necesario tomar en consideración la probabilidad condicional $P(w_a|w_b)$ y la probabilidad $P(w_b)$ del evento que condiciona la observación de dicho evento conjunto: $P(w_a, w_b) = P(w_a|w_b)P(w_b)$. Es de esta forma que Shannon (1948) concibe la entropía de un conjunto de símbolos que son escogidos para construir un mensaje. Con ello, llevar a cabo el cálculo de la entropía implica la generación de estructuras determinadas por la probabilidad de los símbolos de la lengua y de su coocurrencia. Este hecho es objeto de un estudio detallado realizado por Kuich (1970), usando como base a Chomsky and Schützenberger (1963); ideas que a su vez Harris (1957) ya había propuesto.

Los términos de la Eq. (3.2) calculados para un segmento de texto, palabra por palabra, se manifiestan en fluctuaciones de entropía que le permiten al hablante transmitir (o adquirir) información sobre un mensaje a medida que la secuencia de palabras correspondiente es emitida: $w_1, w_2, \ldots, w_i, \cdots \in S_1, S_2, \ldots, S_{N_S}$. La ponderación de cada término de (3.2), por lo tanto, permite conocer qué tan informativo es, en valor esperado con respecto a $P(w_i)$, cada evento conjunto (w_i, S_j). Ello define al vector TF para una oración:

$$\varphi_{ij} = \frac{1}{F}\left(f_{1j}, \ldots, f_{|S_j|j}\right). \tag{3.5}$$

Retomando el modelo de serie ponderada (3.1), se tienen sus coeficientes

$$\begin{aligned} \varphi_{iS_j} &= P(w_i|S_j)\left(\log P(S_j|w_i) - \log P(S_j)\right) \\ &= \frac{f_{ij}}{F}\left(\log \frac{1}{N_{w_i}} - \log \frac{1}{N_S}\right) \\ &= \frac{f_{ij}}{F}\left(\log \frac{N_S}{N_{w_i}}\right). \end{aligned} \tag{3.6}$$

Cada uno de estos proviene de cada componente del producto de Hadamard-Schur $\varphi_j \in \mathbb{R}^{|S_j|}$ entre el vector IDF (3.3) y el vector TF (3.5): $\varphi_j = \varphi_{ij} \odot \varphi_{S_j} =$

$(\varphi_{1S_j}, \ldots, \varphi_{iS_j} \ldots, \varphi_{|S_j|S_j})$.

Las componentes de φ_j entonces regulan la contribución de información que cada representación x_{w_i} de la serie (3.1) aporta a la representación vectorial $s_j(\cdot,\cdot) \in \mathbb{R}^d$ de la oración S_j:

$$\begin{aligned} s_j(x, \varphi_j) &= \sum_{w_i \in S_j} P(w_i|S_j) \log \frac{P(S_j|w_i)}{P(S_j)} \; x_{w_i} \\ &= \sum_{w_i \in S_j} \frac{f_{ij}}{F} \left(\log \frac{N_S}{N_{w_i}} \right) \; x_{w_i} \\ &= \sum_{w_i \in S_j} \varphi_{iS_j} \; x_{w_i}. \end{aligned} \tag{3.7}$$

Nótese que a final de cuentas, los coeficientes están dados por una distribución que probabilidad que pondera a los parámetros x_{w_i} de otra distribución

$$\begin{aligned} P_{x_{w_i}}(w_i|c_k) &= P_{x_{w_i}}(w_i|w_{i\pm1}, \ldots, w_{i\pm|c_k|/2}) \\ &= \frac{1}{Z} exp(x_{w_i}^\top x_{c_k}), \end{aligned} \tag{3.8}$$

donde

$$x_{w_i} = \arg\max_{x_w} \sum_{k=1}^{K} \log P(w_i|x_{c_k}, x_w) \tag{3.9}$$

y $x_{c_k} = (x_{w_{i\pm1}} + \cdots + x_{w_{i\pm|c_k|/2}})/|c_k|$. En la Eq. (3.9), el objetivo es seleccionar al parámetro $x_w \in \mathbb{R}^d$, tal que maximice la verosimilitud $\log P(w_i|x_{c_k}, x_w)$ para todos los K contextos c_k que rodean a los tokens w correspondientes a w_i en un corpus de lengua general $D \ni w$. El parámetro $x_w = x_{w_i}$ cumple con esta condición óptima.

Como ejemplo de uso, suponemos que alguno de los modelos descritos en la sección 2.2 ya está entrenado. Por ejemplo, Glove (sección 2.2.2). Por lo tanto, ahora es posible cargar instancias de representaciones $x_{w_i} \in \mathbb{R}^d$ a partir de dicho modelo. Así, en la figura 3.3, una representación se puede esquematizar en el bosquejo de la frase "el perro ladra". El modelo (3.7) permite ver cómo cierta representación $s(\cdot,\cdot)$ se vería geométricamente (figura 3.4). El bosquejo de la representación del ejemplo podría tener coeficientes como estos: $\varphi_{The} = 0.075, \varphi_{dog} = 0.53, \varphi_{barks} = 0.37$. Por lo tanto, el ejemplo $s[(x_{The}, x_{dog}, x_{barks}), \varphi]$ puede ser descompuesto en

$$\begin{aligned} s(x, \varphi) = s[(x_{The}, x_{dog}, x_{barks}), \varphi] &= \sum_{w_i \in s} \langle \varphi_{w_i}^{(s)}, \varphi_S \rangle x_{w_i} \\ &= 0.075 x_{The} + 0.53 x_{dog} + 0.37 x_{barks} \end{aligned}$$

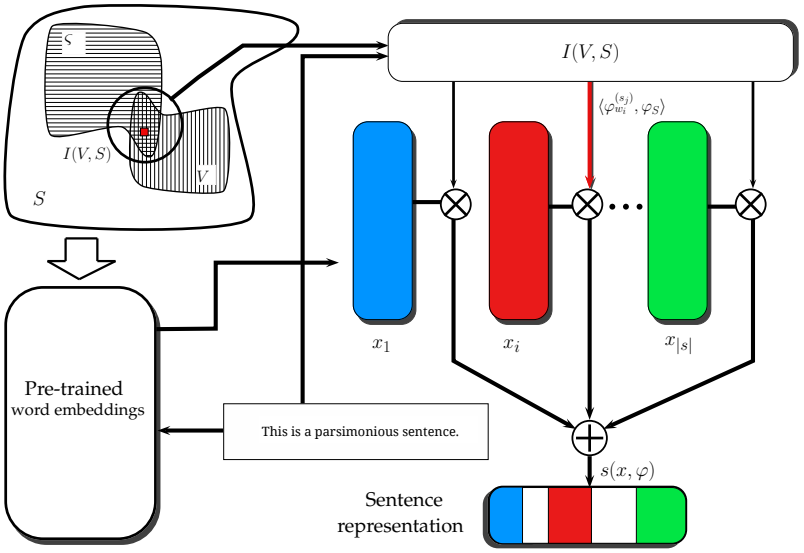

Fig. 3.3: Esquema modular del modelo propuesto.

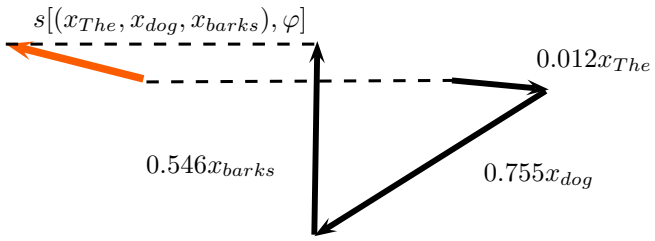

Fig. 3.4: El bosquejo de una oración representada mediante una serie ponderada por entropía de la información.

La modularidad del modelo propuesto permite la incorporación de otras fuentes de información a través de sus coeficientes según sea necesario. Por ejemplo, aunque ya se ha insinuado a lo largo del capítulo que los coeficientes ya tienen una estructura. Se podrían combinar con información adquirida de etiquetas PoS de manera explícita (De Marcken, 1999, Ferrero et al., 2017, Těšitělová, 1992). Además de su ya bien conocido uso como heurísticas para construir bolsas de palabras, los $n-$gramas de palabras o de caracteres constituyen un asunto que requiere mayor atención desde el punto de vista teórico. Pensando en que pueden constituir, sin duda, estructuras de información adicional.

3.3. Costo en tiempo y memoria

Supóngase que se tiene un sistema de NLP que requiere la aplicación del modelo de representación de oraciones propuesto en este trabajo de tesis. Supóngase

también que se tienen representaciones de palabras y sus coeficientes preentrenados (p. ej. TF–IDF). Entonces las representaciones que la aplicación necesita no requieren ser inferidas de forma explícita antes de que sean usadas. Estas pueden ser calculadas en línea (*online*) a medida que dicha aplicación las requiera.

La inferencia en línea de las representaciones es posible gracias a un par de factores sencillos:

- la modularidad del modelo propuesto y
- las operaciones de bajo costo que se requieren para embeber una oración.

Dada una oración de n palabras, WISSE necesita calcular d multiplicaciones escalares por cada palabra de la oración. Es decir, cada dimensión $1, \ldots, d$ de la representación de una palabra será multiplicada por la correspondiente contribución de información $\varphi_i \in \mathbb{R}$, un único coeficiente escalar para todas las dimensiones. Una vez ponderados, WISSE calcula también la suma de las n representaciones. Si se obvian los ciclos de reloj requeridos por los accesos a memoria y asignaciones, entonces el tiempo necesario para embeber una oración es únicamente $T_s = T_{mul}(dn) + T_{sum}(nd) = T_s(2nd)$. Dado que el promedio de longitud por oración es muy corto (por ejemplo, $n \in \{8, 12\}$ palabras para el inglés), se observa que $n \ll d$, de modo que $T_s(2nd) \to \mathcal{O}(kd)$, que es una cota lineal dada por la dimensión de las representaciones, que raramente es mayor a 300.

En cuanto a los requisitos de memoria, los elementos del modelo pueden ser indexados (por ejemplo, en una base de datos o en un directorio del sistema operativo). En este caso, para cada oración solo se cargan en memoria las n representaciones de d dimensiones y sus n coeficientes asociados, es decir, se usan $n + nd = n(1 + d)$ localidades de memoria a la vez. Se tiene en línea una implementación eficiente del método propuesto en este trabajo*.

*`https://github.com/iarroyof/sentence_embedding`

Capítulo 4

Experimentos y resultados

En este capítulo se presentan tanto los múltiples aspectos considerados en los experimentos de evaluación como los resultados obtenidos en dichos experimentos. Primero se describen los conjuntos de datos que se utilizaron para la evaluación de las representaciones de oraciones en tareas de STS. Después, se describe el conjunto de hiperparámetros de las representaciones con los cuales es posible seleccionar un conjunto de representaciones más adecuado para cada una de las diferentes tareas de STS. Como resultado se muestra la correlación entre la similitud anotada por humanos y la similitud obtenida usando representaciones reportadas en el estado del arte, incluyendo las formuladas en este trabajo. Por último, se muestra una comparación global de desempeño entre estas representaciones y las del estado del arte, incluyendo además sistemas de STS (que no requieren representación de oraciones).

4.1. Conjuntos de datos

La evaluación de modelos de representación de oraciones se lleva a cabo comúnmente usando conjuntos de datos creados para diferentes tareas, incluyendo similitud semántica, clasificación de sentimientos, inferencias en lenguaje natural, etc. Este trabajo se concentra en representación de oraciones por su contenido semántico. Por lo tanto se desea evaluar tales representaciones en tareas de medición de similitud semántica (STS). Los conjuntos de datos más reconocidos para ello son SICK y SemEval-2016 (Agirre et al., 2016, Bentivogli et al., 2016, Cer et al., 2017, King et al., 2016). Estos comprenden compilaciones de pares de oraciones procedentes de múltiples fuentes de datos textuales que están abiertas al público. La mayoría de estas oraciones fueron escritas por los usuarios de Internet con estilos de escritura tanto formales como informales; además, las recopilaciones se

hicieron en diferentes años. Una parte importante de las fuentes de datos que usa SemEval son foros y diarios digitales. Parte de los datos también procede de diccionarios digitales y bases de conocimiento (p. ej. Wikipedia, WordNet, OntoNet FrameNet, y otros). Así, los conjuntos de pares de oraciones con los que se evalúa al método de representación de este trabajo son lingüísticamente variados. Esto representa una buena referencia para los resultados obtenidos.

En todos los casos (en SICK y SemEval-2016), los pares de oraciones están asociados a mediciones promedio de similitud anotadas manualmente por humanos (el *gold standard*). La medición va desde 0.0 hasta 5.0 (para pares no relacionados y para pares equivalentes o literalmente idénticos, respectivamente). Los detalles sobre el protocolo de recolección de oraciones y anotación manual de sus similitudes se describen en (Agirre et al., 2016, Bentivogli et al., 2016). Un descripción general de cada conjunto de datos en cuestión se muestra a continuación.

Answer-Answer (SemEval-2016). Este conjunto de datos contiene 1572 pares de respuestas extraídas del repositorio de *StackExchange*. Los pares de respuestas corresponden a foros temáticos relacionados con intercambio académico, cocina, café, viajes, etc.

Headlines (SemEval-2016). Este conjunto de datos contiene 1498 pares de encabezados de noticias, los cuales fueron recopilados por el Observatorio Europeo de Medios (EMM, *European Media Monitor*).

Postediting (SemEval-2016). Este conjunto de datos contiene 3287 pares de oraciones producidas al corregir manualmente traducciones automáticas de noticias (inglés-español-francés). Las traducciones fueron realizadas usando el sistema de traducción automática llamado MOSES*

Plagiarism (SemEval-2016). Este conjunto de datos contiene 1271 respuestas breves a preguntas sobre temas relacionados con computación. Estas respuestas presentan diversos grados de plagio con respecto a artículos de la Wikipedia.

Question-Question (SemEval-2016). Este conjunto de datos contiene 1555 pares de preguntas del repositorio del foro StackExchange. Las preguntas están relacionadas con temas como academia, cocina, café, viajes, etc.

OnWN (SemEval-2013). Este conjunto de datos contiene 561 pares de definiciones de términos procedentes de WordNet y OntoNotes. Cada par de definiciones corresponde a una o a diferentes acepciones de un término con diferentes grados de similitud.

FNWN (SemEval-2013). Este conjunto de datos contiene 189 pares de definiciones de términos procedentes de WordNet y FrameNet. A diferencia de OnWN,

*`http://www.statmt.org/moses`

los pares de FNWN incluyen también definiciones muy cortas comparadas contra otras muy largas. Esto constituye una tarea de medición de similitud llamada *cross-level STS*, la cual implica un alto grado de dificultad para los sistemas automáticos debido a la disparidad en la longitud de las oraciones de un par que se compara. Actualmente, esta tarea se conoce también como Inferencia en lenguaje natural (*Natural Language Inference*), más que como similitud semántica.

SICK (SemEval-2014). Este conjunto de datos contiene 4906 pares de oraciones seleccionadas a partir de conjuntos de datos de años anteriores de la competencia de SemEval (2012-2014). El conjunto de datos está separado en 4 subconjuntos (conjuntos de entrenamiento y prueba. Y dentro de ellos, *entailment* y STS). Tal como se hace en los trabajos del estado del arte, en los experimentos presentados en este trabajo se usaron los conjuntos de entrenamiento y prueba como uno solo, y sin hacer distinción entre *entailment* y STS.

La Wikipedia. Este conjunto de datos contiene 5.11 millones de tipos y fue descargado desde el repositorio de datos de la Wikipedia (2012). Fue utilizado para entrenar a los modelos de representaciones de palabras usados en este trabajo.

4.2. Evaluación de representaciones

La preparación inicial de nuestros experimentos consistió en entrenar representaciones de palabras usando los métodos descritos en la sección 2.2. Tanto W2V* como FastText† se entrenaron con las dimensiones especificadas en la tabla 4.1. Para cada entrenamiento, se usaron los hiperparámetros habituales. Los más importantes de ellos son: la longitud de la ventana de contexto (= 10) y la frecuencia mínima de palabras (= 2). Los modelos se entrenaron usando la Wikipedia en inglés. En el caso de Glove, se utilizaron representaciones previamente entrenadas de 100, 200 y 300 dimensiones, las cuales están disponibles en el sitio web de los autores‡. Para Dep2Vec, se usaron las representaciones de 300 dimensiones previamente entrenadas que están disponibles en el sitio web de los autores§.

Con respecto a los coeficientes del modelo formulado en este trabajo, se calcularon usando una implementación existente de TF–IDF¶. En modo global se usó la

*`https://code.google.com/archive/p/word2vec`

†`https://github.com/facebookresearch/fastText`

‡`https://nlp.stanford.edu/projects/glove`. El entrenamiento de este modelo es sumamente costoso (en disco y memoria), por ello se usaron representaciones previamente entrenados y disponibles en el sitio web del desarrollador.

§`http://u.cs.biu.ac.il/~yogo/data/syntemb/deps.words.bz2`. El procedimiento para entrenar este último modelo no está claramente documentado y por ello se usaron representaciones previamente entrenadas también.

¶`https://radimrehurek.com/gensim/models/lsimodel.html`

Tabla 4.1: (Hiper-)parámetros de WISSE.

Hiperparámetro	Descripción	Valores
Conjunto de datos	Conjunto de datos de evaluación STS	`Plagiarism`, `Answer-Answer`, `Headlines`, `Postediting`, `Question-Question`, `FNWN`, `OnWN` and `SICK` (see Section 4.1)
Embedding	Método de representación	`W2V`, `Glove`, `FastText` and `Dep2Vec`
Dimensión	Dimensiones de las representaciones	`100d`,`200d`, `300d`, `400d`, `500d`, `700d`, `1000d`.
Combinación	Método de combinación de las representaciones	Suma (`sum`), promedio (`avg`).
Coeficientes	Método de ponderación que proporciona los coeficientes de la serie (WISSE). Todos los esquemas de ponderación fueron calculados opcionalmente sin palabras vacías. Para denotarlo, se añadió el sufijo `-st`. De la misma manera, se añadieron los sufijos `-bin` y `-log` para indicar si el vector TF se calculó como binario o como el logaritmo de la frecuencia de las palabras en cada oración.	• TF–IDF global –con Wikipedia (`glob-tfidf`), • IDF global (`glob-idf`), • TF–IDF local –con conjunto STS (`loc-tfidf`), • IDF local (`loc-idf`), • Todos los pesos iguales a `1.0`.
Similitud	Función de similitud para medir STS	`Coseno`, `Euclidiana`, `Manhattan`.

Wikipedia. Para el modo local, la misma implementación fue usada para obtener los coeficientes a partir de cada conjunto de datos STS sobre el cual se hizo la evaluación. Excepto por el tipo de TF (frecuencias, binario o logaritmo), en general los hiperparámetros usados para los coeficientes fueron los predeterminados, salvo aquellos casos que explícitamente se especifican en la tabla 4.1.

Con la finalidad de tener una idea fiable de las posibilidades de WISSE, se llevaron acabo una serie de experimentos utilizando conjuntos de datos enumerados en la sección 4.1. En las evaluaciones hechas usando tanto SICK como SemEval-2016 primero se seleccionaron los hiperparámetros de WISSE. Estos, así como sus efectos sobre el comportamiento del modelo, se resumen en la tabla 4.1. La selección de hiperparámetros fue hecha observando el coeficiente de Pearson entre las similitudes predichas usando las representaciones provistas por WISSE y las similitudes anotadas manualmente por humanos. Para cada conjunto de datos, se seleccionaron los hiperparámetros que dieron el coeficiente de Pearson máximo.

Además de la selección de hiperparámetros de WISSE, se decidió poner a prueba tres diferentes funciones sencillas para calcular la similitud semántica entre pares de representaciones embebidas por WISSE: la similitud coseno, la distancia euclidiana y la distancia de Manhattan. El resultado de estas tres funciones fue utilizado directamente como una predicción de similitud (véase la sección 2.5.2). Para cada conjunto de datos, se compararon los coeficientes de Pearson obtenidos por los métodos del estado del arte contra el mejor obtenido por WISSE.

4.3. Resultados

En esta sección se presentan los resultados de los experimentos de evaluación de las representaciones de oraciones modeladas en esta tesis. Esta evaluación usa métricas conocidas para medir el desempeño de WISSE en tareas de similitud semántica textual para los conjuntos de datos mostrados en la sección 4.1. Estas mediciones de desempeño se dividen principalmente en dos partes. En primer lugar, se presentan los resultados de los experimentos sobre SICK. Esta etapa a su vez está dividida en dos partes. La primera es la selección de hiperparámetros. La segunda es la comparación entre el mejor resultado obtenido con WISSE y los mejores resultados obtenidos por el estado del arte en modelos no supervisados de representación de oraciones.

En segundo lugar, y como en el caso de SICK, en una primera etapa se seleccionaron los hiperparámetros de WISSE para para cada uno de los conjuntos de datos de SemEval-2016 (sección 4.1). En la segunda etapa, los mejores resultados fueron comparados contra el estado del arte.

Por último, se presenta una comparación estadística entre los mejores resultados obtenidos con WISSE y los mejores resultados obtenidos por el estado del arte, tanto de métodos de representación de oraciones (supervisados y no supervisados) como de sistemas STS en SemEval-2016.

4.3.1. Selección de hiperparámetros para SICK

La combinación de todos los hiperparámetros de WISSE resultó en más de 300 experimentos. Sin embargo, en la tabla 4.2 se muestran sólo las diez mejores combinaciones de hiperparámetros, así como las correspondientes medidas de similitud calculadas usando tres funciones, i.e. coseno, Manhattan y euclidiana.

Tabla 4.2: Combinación de hiperparámetros de WISSE. Resultados de correlación con humanos sobre SICK.

Coeficientes	Comb.	Dim.	Embedding	Coseno ρ	Euclid. ρ	Manhatt. ρ
glob-tfidf-bin-st	sum	300d	FastText	**0.72405**	0.64465	0.64387
glob-tfidf	sum	200d	FastText	0.72023	0.64657	0.6469
glob-tfidf-bin	sum	300d	W2V	0.71995	0.66747	0.66751
loc-tfidf-log	sum	350d	FastText	0.71905	0.65583	0.65615
loc-tfidf-bin	sum	400d	FastText	0.71852	0.65236	0.65209
loc-tfidf-st	avg	300d	Glove	0.70397	0.61817	0.61885
glob-tfidf-st	sum	300d	Dep2Vec	0.67925	0.60972	0.61018
loc-tfidf-log	avg	300d	W2V	0.67428	0.6199	0.62075
loc-tfidf-bin-st	avg	300d	W2V	0.66410	0.58308	0.58289
loc-tfidf-log	avg	300d	Dep2Vec	0.64762	0.55620	0.55585

WISSE alcanzó una correlación máxima de $\rho = 0.72405$ con respecto a humanos usando la similitud coseno. Los hiperparámetros usados para alcanzar este resultado se describe a continuación: los coeficientes IDF fueron entrenados en modo global (`glob`) usando la Wikipedia. Los vectores TF fueron calculados como binarios (`bin`) a partir de las palabras en cada oración de SICK, es decir, $\varphi_{w_{ji}} = \{f_{ij} > 0 \ ? \ 1.0/F : 0.0\}$. Además, se suprimieron las palabras vacías (`-st`) y, por consiguiente, fueron omitidas también del embebido de oraciones. Se usaron representaciones FastText de 300 dimensiones (`300d`). Una vez que los representaciones fueron ponderadas se combinaron por sumatoria (`sum`). Para el conjunto de datos de esta sección, FasText resultó mejor que W2V, Glove y dep2Vec.

También se hicieron experimentos usando el promedio de las representaciones `avg`, tanto ponderados como sin ponderar. Cabe señalar que esta operación de combinación es probablemente la única que se usa para generar representaciones de ventanas de contexto en todos los métodos de representación de palabras. Los resultados presentados en este capítulo mostraron que representar una oración como el promedio de las representaciones de las palabras que la forman no ofrece mejores resultados en ningún caso.

Se observa que para SICK, ninguna otra función de similitud fue mejor que coseno. Por lo tanto, este conjunto de datos parece estar mejor caracterizado por los ángulos entre los representaciones que componen una oración.

4.3.2. Resultados para SICK

El principal resultado que se presenta en esta sección es el rendimiento de WISSE sobre el conjunto de datos SICK en comparación con el estado del arte. Las correlaciones entre las anotaciones manuales y las predichas por estos métodos se muestran en la tabla 4.3.

Hasta donde se sabe, al momento de escritura de este trabajo de tesis, tanto Arora et al. (2017) como Pagliardini et al. (2017) proponen los mejores métodos de representación no supervisada de oraciones. Los otros métodos mostrados en la tabla 4.3 son también no supervisados y se consideran como del estado del arte. La línea de base consta de representaciones dispersas TF–IDF con TF binario. Al observar el coeficiente de correlación con respecto a las anotaciones manuales, WISSE tuvo el mayor rendimiento entre los métodos del estado del arte ($\rho = 0.724$). Esto incluye a los que habían registrado los mayores rendimientos hasta ahora. Es importante notar que la diferencia no es muy grande entre WISSE y los dos métodos siguientes en la clasificación (tabla 4.3). De hecho, puede observarse que es muy difícil superar la barrera de $\rho > 0.7200$. Véase que además de WISSE hay 4 métodos que apenas la superan (Glove+WR, Sent2vec, FastSent, C-PHRASE

Tabla 4.3: Rendimiento de métodos no supervisados de representación sobre SICK. Los números en negrita indican el mejor desempeño.

Método	Pearson
Glove+WR (Arora et al., 2017)	0.722
Sent2vec (Pagliardini et al., 2017)	0.720
FastSent (Hill et al., 2016)	0.720
C-PHRASE (Pham et al., 2015)	0.720
CHARAGRAM-PHRASE (Wieting et al., 2016)	0.700
Skip-thoughts (Kiros et al., 2015)	0.600
BoW TF-IDF (Salton et al., 1983)	0.580
SDAE (Hill et al., 2016)	0.460
Doc2Vec (Le and Mikolov, 2014)	0.460
SAE (Hill et al., 2016)	0.310
WISSE (`glob-tfidf-bin-st,sum,300d,FastText`)	**0.724**

y WISSE). Estos están completamente basados en redes neuronales. Con WISSE no es el caso. Además de una red neuronal (para aprender word embeddings), este método incorpora medidas de entropía para la ponderación de los embeddings.

4.3.3. Selección de hiperparámetros para SemEval

En esta subsección se describe los experimentos realizados para la selección de los hiperparámetros de WISSE sobre los conjuntos de datos de SemEval-2016 (véase el tabla 4.4). Estos conjuntos presentan mucho más variedad tanto lingüística como en la forma del texto en comparación con SICK. Entonces, tanto los resultados como la naturaleza de los hiperparámetros también son variados.

Tabla 4.4: Combinación de hiperparámetros de WISSE para la tarea de SemEval-2016.

Conjunto STS	Embedding	Dim.	Coeficientes	Coseno	Euclid.	Manhatt.
Postediting	Dep2Vec	300d	loc-tfidf	0.652880	**0.821610**	0.819890
Plagiarism	W2V	300d	glob-tfidf-bin	0.775750	**0.806070**	0.805280
Ques.-Ques.	FastText	300d	glob-tfidf	**0.704010**	0.683410	0.681400
Headlines	FastText	200d	glob-tfidf	0.676300	**0.701020**	0.700720
Ans.-Ans.	W2V	1000d	loc-tfidf-log	0.507660	**0.655600**	0.652110
OnWN	W2V	1000d	loc-tfidf-log-st	**0.833070**	0.739430	0.738610
FNWN	FastText	200d	glob-tfidf	**0.458560**	0.350300	0.363520

Para todos los casos la mejor forma de combinar los embeddings de la serie fue la suma (`sum`).

El mejor resultado se dio para el dataset Postediting ($\rho = 0.82161$). En este caso se calcularon los pesos de la sumatoria del modelo en modo local y utilizando frecuencias como TFs. Las palabras vacías resultaron de ayuda, de manera se incluyeron en la sumatoria. Para este conjunto de datos, los mejores embeddings

fueron los de Dep2Vec de 300 dimensiones. La mejor función de similitud fue la distancia euclidiana. Con respecto a esto último, se observó que hay una diferencia relativamente grande entre las métricas (euclidiano y Manhattan) y la similitud coseno. Desde luego, esto muestra que no siempre esta última es la mejor opción para medir similitud, lo que dependerá mucho de la naturaleza o de la geometría de las representaciones. Postediting fue el único conjunto de datos para el cual Dep2Vec resultó ser el mejor. Cabe mencionar que este resultado no es del todo general, pues los embeddings proporcionados por los autores de este método apenas abarcan un vocabulario de 149546 tipos*, lo cual difiere mucho de lo que se puede lograr entrenando, por ejemplo, fastText sobre la Wikipedia (2003634 tipos).

El segundo mejor resultado obtenido por WISSE fue para el conjunto de datos Plagiarism ($\rho = 0.80607$). En este caso, los mejores hiperparámetros variaron considerablemente con respecto a los mejores obtenidos para Postediting. Las principales diferencias se observaron en la forma en que se ajustaron los pesos de la sumatoria. Esto se hizo en modo global y utilizando TFs binarios. Los mejores word embeddings para el conjunto de datos en cuestión fueron los de W2V de 300 dimensiones.

Los resultados más bajos en SemEval-2016 se obtuvieron en el conjunto de datos Answer-Answer ($\rho = 0.6556$). Este ha sido reportado como un reto significativo para la mayoría de los métodos/sistemas de la competencia STS (Agirre et al., 2016). Para este conjunto de datos, WISSE se desempeñó mejor usando word embeddings W2V de 1000 dimensiones. Los pesos de la serie fueron calculados en modo local y con TFs sublineales; esto es, el logaritmo de las frecuencias de las palabras dentro de cada oración: $\varphi_{w_i}^{(s_j)} = log(f_{ij} + 1)/F$.

En la mayoría de los casos, las dos métricas (euclidiana y Manhattan) superaron a la similitud coseno y la diferencia es relativamente grande. Se observó esto para los conjuntos de datos Postediting, Plagiarism, Headlines y Answer-Answer. De igual forma, la sumatoria fue la mejor operación de combinación de embeddings ponderados para construir las representaciones de oraciones.

En general la dimensionalidad de los embeddings mostró algunas regularidades. En la mayoría de los casos los resultados fueron mejores usando 300 dimensiones. Asimismo, los casos de OnWN y Answer-Answer son interesantes, ya que se requirieron embeddings de 1000 dimensiones para alcanzar correlaciones apenas aceptables, lo que aparentemente también está relacionado con que los TFs fuesen sublineales (pero, ¿porqué se requieren TFs sublineales para ponderar vectores densos de mayor dimension? ¿De qué naturaleza son las oraciones de es-

*Una muy posible razón del vocabulario reducido de este método es la necesidad de calcular vectores de contexto a partir del árbol de pendencias. Este último paso es computacionalmente muy costoso y es muy posible que limite en gran medida la representación distribuida original de W2V.

tos conjuntos de datos, tal que se requieren 1000 dimensiones para representarlas mejor?). Otro detalle interesante de estos conjuntos de datos es que permitieron variar la dimensionalidad de los embeddings de manera incremental en el rango $\{200, 1000\}$. No obstante, una vez que los incrementos superaron las 300 dimensiones el rendimiento casi no mejoró; lo hizo muy lentamente. En contraste, para todos los demás conjuntos de datos produjo un efecto diferente. Aumentar las dimensiones en este mismo rango resultó en correlaciones máximas para aproximadamente 300 dimensiones, pero al sobrepasar dicho valor el rendimiento decayó.

En cuanto a sus coeficientes, las representaciones de oraciones se desempeñaron mejor en todos los experimentos cuando estos fueron productos TF–IDF (representados como $\langle \varphi_{w_i}^{(s_j)}, \varphi_S \rangle$ en nuestro modelo). Al igual que en el caso del promedio de word embeddings, un experimento típico en la literatura es usar sólo los IDF $\langle \mathbb{1}, \varphi_S \rangle$ como coeficientes de una sumatoria para representar oraciones. Los experimentos realizados con este método mostraron bajos resultados en general. Por lo tanto, este comportamiento puede considerarse independiente de todos los hiperparámetros del modelo, incluyendo de método de word embedding y sus dimensiones.

Originalmente en esta tesis se pensó en dedicar nuestro modelo para representar definiciones de términos. Entonces se hicieron experimentos de similitud semántica con OnWN y FNWN (SemEval-2013), que son dos conjuntos de datos que contienen definiciones de términos. Estos conjuntos, de manera similar a Answer-Answer, han sido reportados como retadores para los métodos y sistemas del estado del arte (Agirre et al., 2013).

Las representaciones dispersas TF–IDF de la línea de base han mostrado resultados satisfactorios para OnWN ($\rho = 0.8431$). Esto, proporciona un punto de comparación difícil de superar, pensando además en que se trata de un método relativamente sencillo. Pero por otro lado, este tipo de representaciones no ha mostrado estabilidad en diversos escenarios, lo que lo hace menos viable en aplicaciones de propósito general. Esto se puede observar en las tablas 4.3 y 4.5, donde se observa como esta línea de base presenta un desempeño bajo en la mayoría de las tareas de STS. El mejor resultado de WISSE sobre OnWN fue comparable al de TF–IDF ($\rho = 0.82161$) usando embeddings W2V de 1000 dimensiones. Los vectores IDF fueron ajustados en modo local y los TFs fueron sublineales. Las palabras vacías fueron suprimidas de las representaciones. Con FNWN, los sistemas competidores (sin restricción de tipos de métodos en cuanto supervisión y uso de recursos externos) apenas han alcanzado una correlación máxima de $\rho = 0.5818$. WISSE alcanzó $\rho = 0.45856$ en este conjunto de datos utilizando embeddings FastText de 200 dimensiones. Los pesos de la serie fueron ajustados en modo global (IDF)

y con TFs de frecuencias. Para ambos conjuntos con definiciones de términos, la similitud coseno dio resultados considerablemente mejores que las métricas.

4.3.4. Resultados para SemEval

En esta subsección se presentan los resultados de comparación entre WISSE y los métodos no supervisados del estado del arte usando los conjuntos de datos de SemEval-2016 (tabla 4.5). Para estos se han usado los mejores hiperparámetros de WISSE para cada conjunto de datos (sección 4.1). Se observó que tanto para Plagiarism como para Answer-Answer, WISSE superó al estado del arte. Para los otros tres conjuntos de datos (Headlines [HDL], Postediting y Question-Question) Sent2vec fue mejor. Se observó que, en general, para Postediting es difícil de superar el desempeño de la línea de base (representaciones dispersas TF–IDF). De hecho, Sent2Vec lo sobrepasó por una muy pequeña diferencia. Asimismo WISSE estuvo debajo, pero también, por una muy pequeña diferencia.

Tabla 4.5: Evaluación de WISSE y el estado del arte de modelos no supervisados sobre SemEval-2016

Modelo	Ans.-Ans.	HDL	Plagiarism	Postediting	Ques.-Ques.
Sent2Vec*	0.57739	***0.75061***	0.80068	***0.82857***	***0.73035***
WISSE	**0.65560**	0.70102	**0.80607**	0.82161	0.70410
D2V (400d)	0.41123	0.69169	0.60488	0.75547	-0.02245
Skip-toughs	0.23199	0.49643	0.48636	0.17749	0.33446
W2V (300d-avg)	0.50311	0.66362	0.72347	0.73935	0.16586
BoW (TF binario)	0.41133	0.54073	0.69601	0.82615	0.03844

*Este método fue evaluado por el equipo MayoNLP (Afzal et al., 2016).

A pesar de que actualmente D2V (*ParagraphVector*) y Skip-Thoughts son métodos muy populares, sus desempeños en tareas de STS son demasiado bajos (incluso no se acercan a la línea de base en Postediting). Los desempeños de estos métodos fueron considerablemente inferiores a los de Sent2Vec y WISSE (que se encuentran en los lugares más altos de la clasificación). Por ejemplo, mientras que Sent2vec y WISSE alcanzaron $\rho > 0.70$ en Question-Question, D2V cayó hasta $\rho = 0.02245 < 0.1$. Una situación similar ocurrió con Skip-Thoughts, el cual sólo alcanza $\rho = 0.17749$ en Postediting (véase que los mejores métodos ofrecen $\rho > 0.82$). Por último, véase que D2V apenas se compara con el promedio simple de embeddings W2V para construir representaciones de oraciones (Skip-Thoughts está incluso más debajo). Esto hace completamente inviable embeber representaciones de oraciones usando D2V o Skip-Thoughts para tareas de STS, puesto que incluso el promedio de embeddings es mucho menos costoso y conceptualmente mucho más simple que ellos.

4.3.5. Posición de WISSE en SemEval

En esta subsección se presentan los resultados de comparación entre WISSE y los diez mejores sistemas participantes en SemEval-2016 (tabla 4.6). Aunque WISSE no participó en esta competencia, es relevante observar cómo se desempeña con respecto al estado del arte de sistemas STS. Nótese que estos sistemas están diseñados en diversas formas y para distintos propósitos que WISSE (sección 2.5.1). Es decir, mientras que WISSE pretende representar oraciones en espacios vectoriales y de manera no supervisada, los sistemas STS presentados en esta comparación están principalmente diseñados para medir la similitud semántica de manera supervisada y sin restricciones de uso de recursos de conocimiento externo o lingüísticos.

Tabla 4.6: Clasificación de desempeños para la tarea de SemEval-2016 (esta tabla es una versión modificada de aquella proporcionada por Agirre et al. (2016).

R	Sistema/método	ALL	Ans.-Ans.	HDL	Plagiarism	Postediting	Ques.-Ques.
1	Samsung Pol.	0.77807	0.69235	0.82749	0.84138	0.83516	0.68705
2	UWB	0.75731	0.62148	0.81886	0.82355	0.82085	0.70199
3	MayoNLPTeam	0.75607	0.61426	0.77263	0.805	0.8484	0.74705
4	Samsung Pol.	0.75468	0.69235	0.82749	0.81288	0.83516	0.58567
5	NaCTeM	0.74865	0.60237	0.8046	0.81478	0.82858	0.69367
6	ECNU	0.75079	0.56979	0.81214	0.82503	0.82342	0.73116
7	UMD-TTIC-UW	0.74201	0.66074	0.79457	0.81541	0.80939	0.61872
9	SimiHawk	0.73774	0.59237	0.81419	0.80566	0.82179	0.65048
8	Sent2Vec	0.73836	0.57739	0.75061	0.80068	0.82857	0.73035
10	**WISSE**	0.73768	0.655600	0.70102	0.80607	0.82065	0.70410
23	UWB	0.72622	0.64442	0.79352	0.82742	0.81209	0.53383
–	D2V (400d)	0.50206	0.41123	0.69169	0.60488	0.75547	-0.02245
–	Skip-toughs	0.27148	0.23199	0.49643	0.48636	0.17749	0.33446
–	W2V (300d-avg)	0.56007	0.50311	0.66362	0.72347	0.73935	0.16586
110	STS (BoW)	0.51334	0.41133	0.54073	0.69601	0.82615	0.03844

Para la comparación de esta subsección se han utilizado los mejores hiperparámetros de WISSE para cada conjunto de datos (sección 4.1). La medida de desempeño general de la competencia es un promedio ponderado de las correlaciones sobre todos los conjuntos de datos de la tarea (columna `All`). Tomando esta medida de desempeño, se observa que WISSE, está en el puesto número 10, entre 113 sistemas que participaron esta competencia. Se considera relevante el hecho de que esta comparación se realizó dejando de lado las características de todos los sistemas o métodos del estado del arte. Para mantener la distinción entre estos sistemas y los sistemas de representación no supervisada se ha colocado una línea a la mitad de la tabla 4.6.

Para confirmar que WISSE es competitivo, se analizaron estadísticas sencillas del coeficiente de correlación obtenido por los sistemas/métodos participantes (figura 4.1 y tabla 4.6). La primera observación que emerge es que WISSE (denotado

por un diamante verde en la figura) superó a la media del desempeño en la mayoría de los conjuntos de datos. En el caso de Headlines, WISSE no superó la media, pero se mantuvo en el rango intercuartil (IQR). En general, tanto Plagiarism como Postediting no presentan dificultades para el estado del arte y la línea de base (TF–IDF, círculo rojo). Esto puede leerse en la alta correlación media alcanzada por la mayoría de los sistemas y su variabilidad baja. De hecho, la línea de base superó la media. Para estos conjuntos de datos, WISSE se mantiene dentro del IQR, pero no demasiado lejos de la correlación máxima.

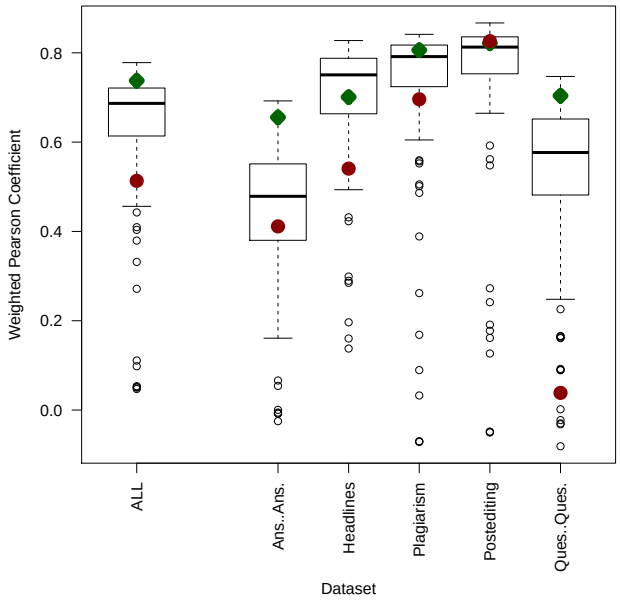

Fig. 4.1: Diagrama de caja de la situación estadística de WISSE y el estado del arte en la competencia SemEval-2016. El eje vertical es el coeficiente de correlación. El eje horizontal indica cada conjunto de datos. Marcadores: WISSE = diamante verde; línea de base = círculo rojo.

Estadísticamente, los mejores resultados de WISSE fueron alcanzados tanto para Question-Question como para Answer-Answer. WISSE se posicionó más allá del primer cuartil, cerca del máximo general. Es decir, WISSE superó la incertidumbre de las correlaciones alcanzadas por el 95 % de los sistemas (IQR). Para Question-Question la línea de base se encuentra en la zona de correlaciones atípicas bajas. Nótese que Answer-Answer es uno de los conjuntos de datos más difíciles al predecir similitudes para el estado del arte, pero aún así la línea de base alcanza el IQR.

Para el desempeño general de correlación (`All`), WISSE se posicionó más allá del primer cuartil; de hecho en el bigote superior del correspondiente diagrama

de caja.

4.4. Aplicación: agrupamiento de contextos de acepción

En este trabajo de considera una aplicación de WISSE en agrupamiento (*clustering*) de segmentos de texto. Estos segmentos de texto tienen la particularidad de que pueden ser definidos como *Contextos ricos en conocimiento*, CRCs (Davidson, 1998, Meyer, 2001), o de manera más general, como *contextos de acepción* (Pedersen et al., 2005, Williamson, 1997). Se ha usado WISSE para representar el contenido de los segmentos (mas no estrictamente su semántica) en un espacio vectorial de dimensión finita. Sin embargo, esta tarea de agrupamiento generalmente se conoce como "agrupamiento semántico" y la llamaremos así por sencillez.

El conjunto de segmentos de texto se recopiló mediante la herramienta de extracción de información (IE) llamada Describe (Sierra, 2009).* Este sistema recibe como entrada una petición, que puede ser una palabra o una frase (frase nominal). Su misión es acceder a muchas páginas web para buscar segmentos de texto en ellas. Estos segmentos deben contener la palabra o frase de la petición. Cada segmento cumple con las reglas de una gramática especificada por Alarcón et al. (2007), la cual sirve como filtro al sistema para presentar los segmentos de texto extraídos. Esta gramática tiene la intención de que los segmentos contengan oraciones lo más similares posible a la definición de la palabra o frase que el sistema recibió como petición. Por lo tanto, en este experimento de aplicación se espera como resultado que, si al menos la mayoría de los segmentos extraídos por el sistema de IE contienen definiciones o al menos CRCs, los segmentos estén agrupados de acuerdo con las diferentes acepciones de la definición de la palabra o frase que recibe como petición el sistema (y posiblemente algunas variantes diferentes de cada acepción).

Algunos ejemplos de extracciones obtenidas con el sistema Describe se encuentran disponibles en línea (Arroyo-Fernández, 2016).† Los segmentos de texto que se analizan en este caso de aplicación son el resultado de hacer una consulta sobre la palabra "tree" (árbol). En la figura 4.2 se muestra el agrupamiento completo para los segmentos de texto extraídos para la palabra mencionada. El agrupamiento se llevó a cabo usando un método muy conocido llamado *Ward's linkage* (Ward Jr, 1963). Este puede operar sobre la matriz de distancias entre los vectores de la matriz de representaciones de segmentos obtenidas mediante WISSE. Se usó la distancia coseno, la cual resultó ser más consistente con las tareas de simili-

*`www.describe.com.mx`

†`http://github.com/iarroyof/describe_corpus`

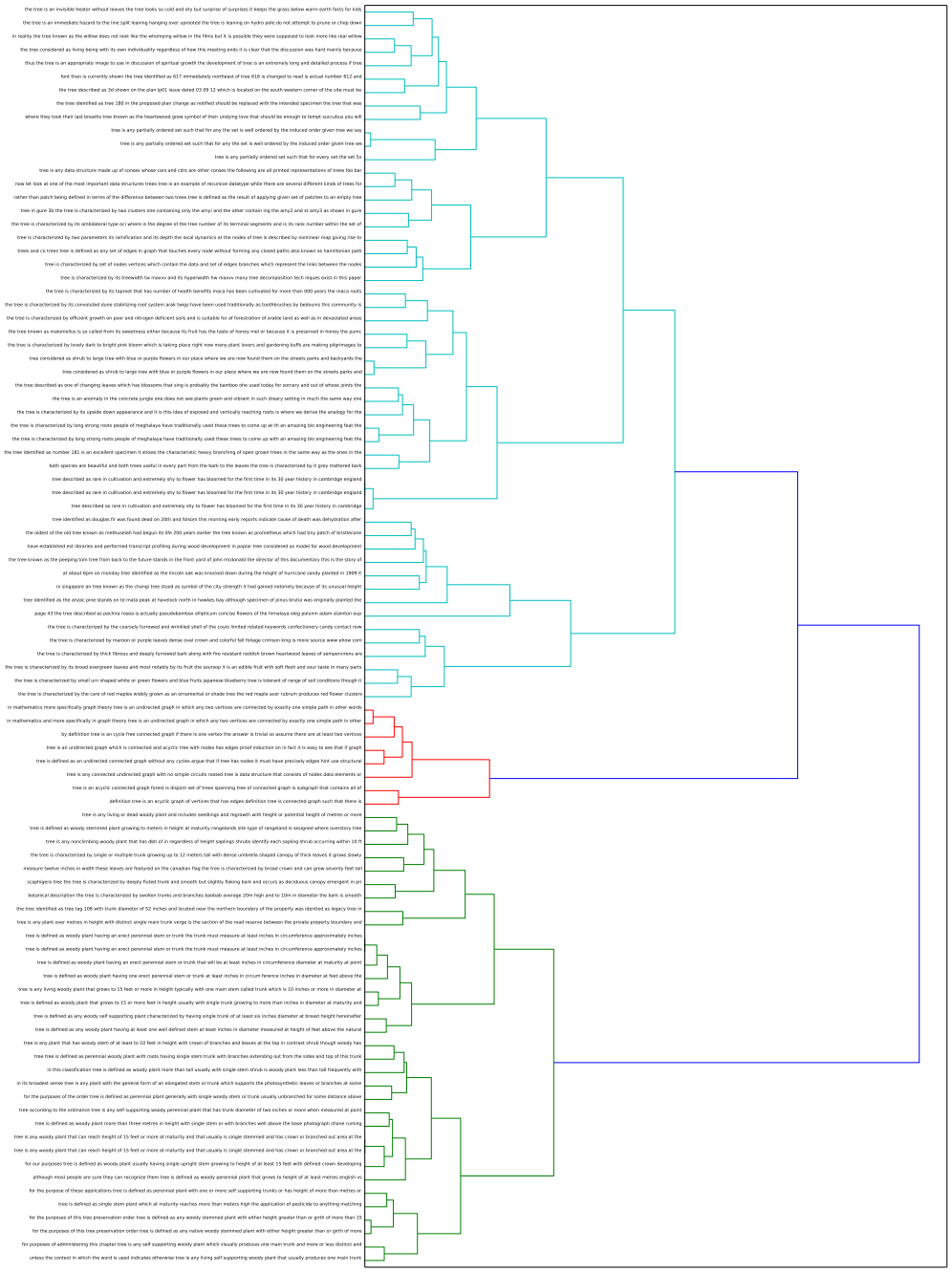

Fig. 4.2: Dendrograma de agrupamiento completo para los segmentos extraídos para la palabra *"Tree"*.

tud semántica realizadas usando representaciones WISSE en SICK. Puesto que no existen anotaciones manuales para la similitud o agrupamiento de los segmentos de texto que se analizan en esta sección, se hará un análisis cualitativo del dendrograma obtenido. Por claridad, este análisis se enfoca en las diferencias más claras

entre las temáticas que detectó el algoritmo de agrupamiento.

En la figura 4.3 se muestra un corte del dendrograma de la figura 4.2. En este corte se tiene un cluster interesante que contiene tres segmentos de texto que hablan del concepto de "árbol" desde el punto de vista de *teoría de conjuntos*. Aunque los tres segmentos son bastante parecidos entre sí, el hecho de que hayan resultado ser muy similares a pesar de que el más corto de ellos contenga a penas la mitad de las palabras que contiene el más grande. Esta diferencia de longitud se refleja en la altura de la parte del dendrograma que une a los dos segmentos más grandes con el más pequeño.

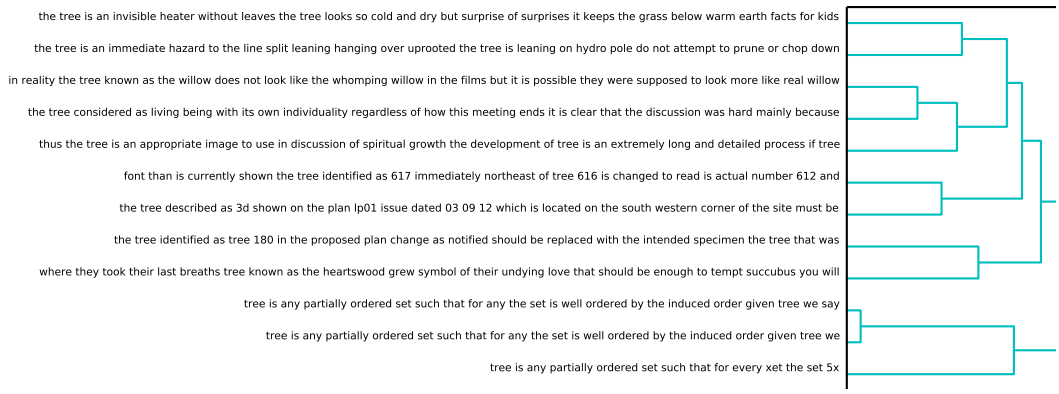

Fig. 4.3: Corte del dendrograma de agrupamiento para la palabra *"Tree"*. Grupo de teoría de conjuntos (primeros tres segmentos, de abajo para arriba).

En la figura 4.4 se muestra un corte del dendrograma de la figura 4.2. En este corte se tiene un cluster interesante que contiene tres segmentos de texto que hablan del concepto de "árbol" desde el punto de vista de *estructuras de datos y teoría de redes*. Esta vez las diferencias son poco evidentes si se toman en cuenta solo elementos léxicos, ya que en este grupo se habla del concepto de árbol como estructura de datos, en niveles de abstracción diferentes y con diferente vocabulario. De hecho, resulta también notable el hecho de que dentro del mismo grupo de computación y estructuras de datos, se puede identificar un grupo de teoría de redes, el cual se puede identificar porque contiene varias palabras relacionadas como *"paths"*, *"edges"*, *"nodes"* (primeros cuatro segmentos). Nótese que la diferencia entre este subgrupo y el grupo de matemáticas puede ser un tanto sutil, pero las representaciones parecen tener suficiente nivel de detalle en el espacio vectorial donde habitan como para que el algoritmo de agrupamiento detectara las diferencias de similitud.

En la figura 4.5 se tiene un grupo de segmentos que en su mayoría hablan de manera más clara sobre *matemáticas*; en particular, de teoría de grafos (o gráfi-

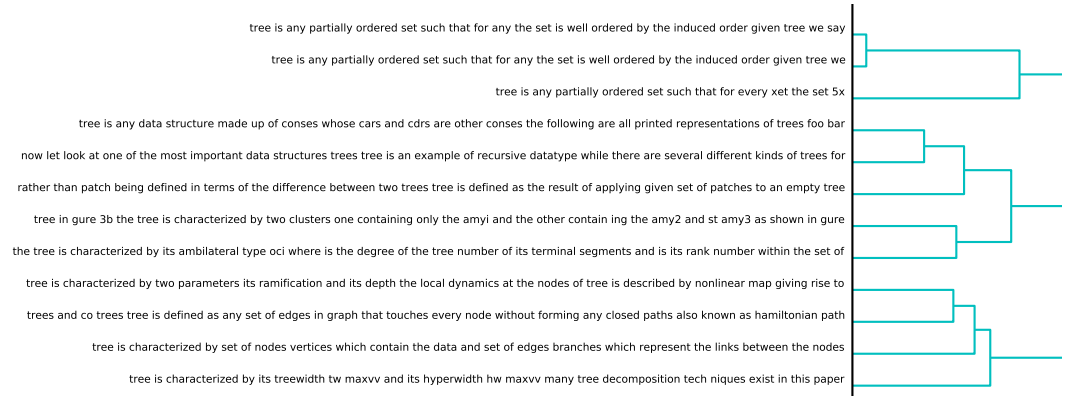

Fig. 4.4: Corte del dendrograma de agrupamiento para la palabra *"Tree"*. Grupo de estructuras de datos y teoría de redes (primer cluster, primeros nueve segmentos, de abajo para arriba).

cas). Nótese que este grupo puede considerarse muy similar al grupo de teoría de redes, ya que algunas palabras, como "path" y "edge" son propias de ambos dominios. Sin embargo, se considera que las representaciones se mantienen un tanto alejadas debido a que la palabra "graph" sólo puede ser observada en el grupo de teoría de grafos, y no en el de teoría de redes. Este hecho sugiere como coeficiente discriminador de valor alto al de esta palabra, haciéndola altamente informativa para las representaciones. Esto porque el número de segmentos que comparten a $\{"paths", "edges", "nodes", \dots\}$ es relativamente mediano (una consecuencia de la estructura de subconjuntos en el corpus). Así, WISSE asigna coeficientes mucho más bajos a estas palabras, pues ocurren usualmente para describir a un árbol en el sentido matemático. Es de esta manera como WISSE induce cierta resolución al espacio de representación. Nótese que esta puede ser un tanto contraproducente en casos donde no se requiera distinguir similitudes relativamente detalladas.

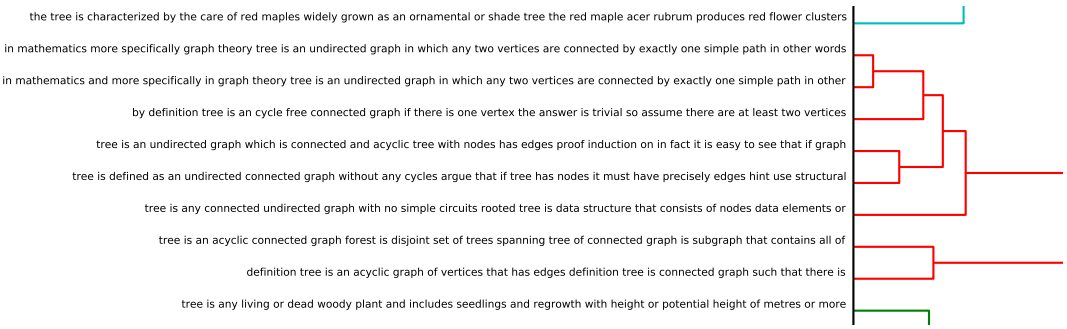

Fig. 4.5: Corte del dendrograma de agrupamiento para la palabra *"Tree"*. Grupo de matemáticas (cluster rojo).

Finalmente en la figura 4.6, se tiene un cluster que contiene segmentos de

texto que hablan del concepto de árbol como *planta* (la acepción principal del la palabra). Se observa que dentro de este cluster existen dos subgrupos. La principal diferencia entre ellos consiste en que el grupo de abajo contiene descripciones más generales sobre un árbol. Estas hablan de sus tres atributos principales: que es de madera, que tiene un tronco y rices; y que crece para tener un tamaño relativamente grande. El grupo de arriba, menciona cuestiones menos regulares pero a la vez más detalladas; tal vez hablando de especies, hábitat, formas de follaje, etc., aunque el contexto no es suficiente para determinarlo con certeza (al menos para el autor de este trabajo). Lo que sí es posible observar es que se mantiene la idea de un árbol como una planta grande.

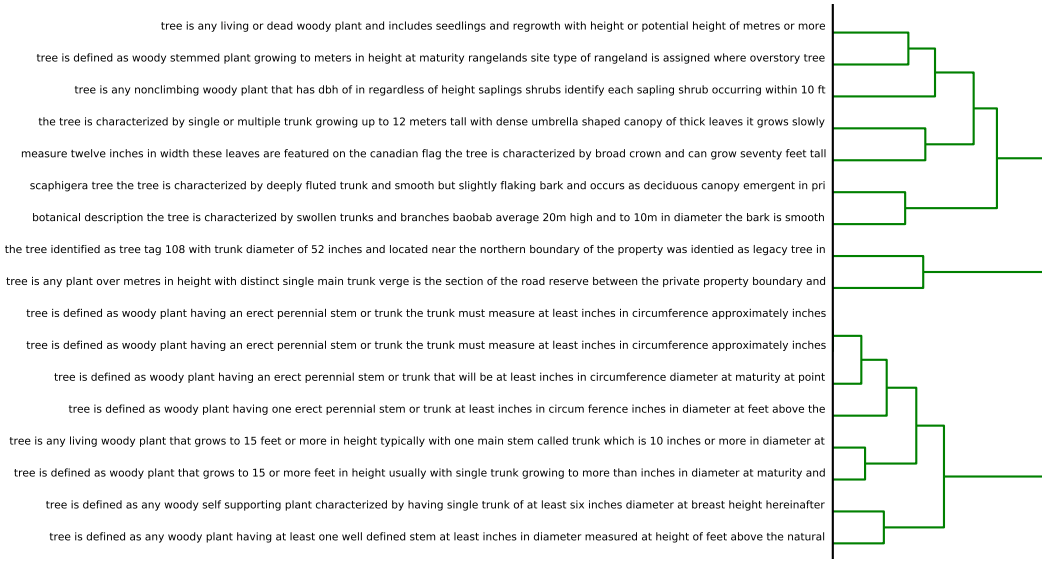

Fig. 4.6: Corte del dendrograma de agrupamiento para la palabra *"Tree"*. Grupo planta (cluster verde).

4.5. Discusión

La entropía de Shannon de las palabras en el corpus puede ser utilizada para ponderar la contribución de información de cada palabra en una oración. Por lo tanto, estos dos elementos (las palabras y sus aportes de información) son identificables al momento en que WISSE construye representaciones de oraciones. Esta identificabilidad puede incentivar estudios futuros sobre las propiedades estadísticas del significado al nivel de oración.

Un aspecto sorprendente del enfoque propuesto es que la entropía es un número real (un escalar). Por lo tanto nuestros experimentos se realizaron relativamente a bajo coste computacional. Esto difiere considerablemente de los enfoques que es-

tán basados completamente en redes neuronales. Por ejemplo, usando LSTMs (o cualquier tipo de RNN) los métodos deben aprender una matriz grande de pesos sinápticos para ponderar cada word embedding y así capturar las interacciones entre ellos con la finalidad de componer una representación de oración. Ello requiere días o semanas de entrenamiento en GPUs. Este no es el caso para D2V, sin embargo los embeddings de las palabras de una oración deben ser proyectados en matrices de pesos sinápticos para predecir una representación de oración.

Se mostró en las secciones 2.5.1 y 2.5.3 que existe una amplia variedad de sistemas de STS. Varios de ellos requieren supervisión y se basan en conocimiento externo y recursos lingüísticos. Como se mostró en la sección 4.3.5, estas ventajas no siempre son significativas en comparación con los métodos no supervisados. En este sentido, se observó que en la mayoría de los casos, WISSE superó la incertidumbre de las correlaciones medias alcanzadas por el 95 % de todos los sistemas del estado del arte (esto es válido aun cuando la variabilidad del rendimiento general es alta). Se considera que los hechos mencionados convierten a WISSE en un método competitivo para representar el significado al nivel de oración.

4.5.1. Propiedades del texto contra desempeño

Las diferencias en el comportamiento de WISSE con respecto a las propiedades textuales de cada conjunto de datos son significativas. Ello demuestra que puede haber propiedades del texto para las cuales este método y los otros son sensibles, produciendo así resultados diferentes. Por lo tanto, el problema de la estabilidad en modelos no supervisados sigue abierto. No obstante, esta variabilidad en las propiedades del texto, permitió observar la flexibilidad de WISSE de ser preparado para diversos escenarios. Es decir, tomando ventaja de la modularidad del modelo es posible utilizar hiperparámetros diferentes de acuerdo con las propiedades del texto.

El hecho de que WISSE superó al estado del arte en SICK es interesante. Esto es debido a que este conjunto de datos constituye una cuidadosa selección de pares de oraciones provenientes de tareas variadas de STS (años 2012-2014). Esta selección incluye tareas del tipo cross-level, que son las más difíciles de superar.

En la tabla 4.7 se muestran la media (μ), la mediana de las longitudes de las oraciones de los conjuntos de datos. Asimismo, se muestran la media y la mediana ($\mu-$dif y $m-$dif, respectivamente) de la diferencia entre las longitudes de los pares de oraciones y los coeficientes de variabilidad, tanto de las longitudes (c_v) como de las diferencias (c_v-dif). Nótese que de las estadísticas mostradas en la tabla, la que más impacto negativo tiene sobre el rendimiento (ρ) de WISSE es la diferencia (cuantificada con $\mu-$dif y $m-$dif) entre las longitudes de las oraciones que se

comparan en un par. En la tabla se puede observar claramente que para FNWN ($\mu-$dif$=21$) WISSE obtuvo cifras bajas de desempeño (la mayoría de los métodos del estado del arte también muestran esta dificultad). Con base en la tabla 4.7 es posible intuir que este bajo desempeño fue debido a la magnitud que las representaciones de oraciones alcanzan al sumar las contribuciones de los embeddings de la sumatoria. Cuanto mayor sea el número de palabras de una oración, mayor será la magnitud alcanzada por la representación resultante*. Entonces, para el caso de FNWN se pide la comparación entre una frase y una oración larga; es decir, un texto de, p. ej., 3 palabras contra otro de, p. ej., 23 palabras (lo que hace una diferencia de 21 palabras en promedio). Esto representa un desequilibrio en las longitudes de los textos que hace que WISSE embeba vectores significativamente diferentes para un par de segmentos de texto. Estos vectores deben ser comparados por una función que en general está definida para geometría euclidiana (coseno, Manhattan y euclidiana). Así, aun cuando los significados representados para un párrafo y una frase sean relativamente similares, el desequilibrio de magnitudes entre sus representaciones puede causar diferencias geométricas significativas. Estas conjeturas se apoyan también con el hecho de que la similitud coseno, que principalmente mide el ángulo entre vectores, fue la mejor para el conjunto de datos en cuestión (sección 4.3.3).

Para el caso de Answer-Answer, no se observa relación entre el desempeño relativamente bajo y el sesgo de la distribución de longitudes o la variabilidad de estas y sus diferencias (medida con el coeficiente de variación c_v). Tampoco la longitud del texto fue un factor. Incluso otros conjuntos de datos presentan mayores sesgos, variabilidades y longitudes medias. La única diferencia con los demás conjuntos es la dimensión de los embeddings que se requirió para representar las oraciones de Answer-Answer.

En general, se observa que cuando existe equilibrio entre las longitudes de los fragmentos de texto a comparar las métricas resultaron mejores. Ello porque estas explotan directamente el hecho de que el significado de la oración fue codificado tanto en el ángulo como en las magnitudes de las representaciones resultantes.

Debe observarse un hecho notable. Las deficiencias que se han descrito arriba son observaciones en tareas de STS, más que en la capacidad de representación de WISSE. Esto se confirma por el hecho de que WISSE no muestra sensibilidad a la longitud de las oraciones a representar, es decir, se puede tener buen desempeño representando oraciones largas de más de 20 palabras y oraciones cortas de 7 palabras en promedio.

*Esto es relativo, ya que para cualquier embedding $x \in \ell_p$ su norma siempre es finita, i.e. $\|x\|_p = (\sum_{i=1}^{\infty} |x_i|^p)^{\frac{1}{p}} < \infty$.

Tabla 4.7: Estadísticas de longitud de texto en los conjuntos de datos y su relación con el desempeño de WISSE y la medida de similitud (Dist.)

Dataset	μ	Mediana	len. c_v	$\mu-$dif	$m-$dif	c_v-dif	ρ	Dist.	Dim.
OnWN	7	7	0.774	1	1	1.02	0.833	Cos	1000
Postediting	20	17	0.296	1	1	1.16	0.821	Eucl.	300
Plagiarism	13	14	0.261	2	2	0.848	0.806	Eucl.	300
SICK	9	9	0.382	1	1	1.16	0.724	Eucl.	300
Ques.-Ques.	10	10	0.310	2	2	0.964	0.704	Cos	300
Headlines	7	8	0.275	1	1	0.936	0.701	Eucl.	200
Ans.-Ans.	9	8	0.434	2	2	0.882	0.655	Eucl.	1000
FNWN	19	14	0.774	21	20	0.732	0.458	Cos	200

4.5.2. ¿Semántica distribucional?

En lingüística existen diversas hipótesis sobre el rol de la semántica en el estudio de las lenguas naturales. Chomsky, por ejemplo, sugiere que si la lingüística es la ciencia que estudia el fenómeno de las lenguas naturales, entonces la semántica no es parte de ella. Esto porque existen razones para pensar que la semántica esta está fuera del fenómeno observado. Esta es una postura interesante, porque, a primera vista, es fácil pensar que si se usa un lenguaje para transmitir un mensaje, es natural que dicho mensaje sea transportando o esté codificado en las emisiones del mismo. Esta última es de hecho la postura en la que muchas teorías se apoyan para considerar que la semántica en efecto es parte de la lingüística.

Como en muchos planteamientos científicos o como en el tratado de cualquier pregunta de investigación, una postura en particular sobre los alcances de la lingüística a fin de cuentas se puede ver como una forma de delimitar el estudio de un problema. Es decir, puede no ser cierto que la semántica no es objeto de estudio de la lingüística; sin embargo, asumir lo contrario puede ser solo una manera de facilitar un análisis dado. Se convierte solo en un punto de vista conveniente para tratar un problema específico (p. ej. en álgebra básica, multiplicamos por -1 una ecuación para facilitar el análisis pero sin afectar el resultado).

Independientemente de cualquier postura, los modelos que se han presentado hasta el momento en este trabajo no han requerido de un estudio exhaustivo de semántica. Esto se mantiene incluso cuando se trata de definir aspectos de la medición de "similitud semántica". Sin duda alguna, es muy posible que lo dicho en el párrafo anterior sea un reflejo de las limitaciones del autor de este trabajo para conocer todas las posturas posibles sobre semántica y estructura. Sin embargo, lo que parece evidente es que los métodos estadísticos usados hasta el momento no parecen haber tenido contacto directo con las ideas que se alojan en las mentes de los hablantes que produjeron los datos de entrenamiento y evaluación. Incluso al final de la sección 4.3, se observa que estadísticamente no son significativas las

mejoras que se pueden alcanzar si se usan tantos recursos lingüísticos y de conocimiento como los haya disponibles para medir la similitud semántica. Todo esto, desde el punto de vista particular del autor de esta tesis, parece indicar que hay evidencias que favorecen las posturas estructuralistas. Como interpretación de estas, se puede decir que la lengua es un conjunto de fluctuaciones que aprovechan la economía de lo informativo y que principalmente transportan mensajes codificados. Esta economía quiere decir que, de manera natural, la información tiende a ser comprimida en forma de un código óptimo (Fano, 1949, Huffman, 1952), tal como lo propone el propio Shannon (1948). Cuando este código es percibido por el cerebro del hablante, este direcciona a la combinación de referentes que ya tiene precargados en su memoria y que en momentos pasados almacenó (Allenby and Rossi, 2006, Bruni et al., 2014, Jenatton et al., 2010a,b). Entonces, estos códigos no son los significados en sí; más bien lo son aquello a lo que apuntan, en la memoria.

¿Qué sí ha sido necesario? contar las intersecciones de (y entre) cadenas terminales en varios niveles jerárquicos para estimar distribuciones de probabilidad de las mismas. El resultado de ello: un vector inferido por un estimador de la media de estas distribuciones. Entonces ¿porqué los resultados de similitud semántica son coherentes con lo que los anotadores humanos midieron? Es por ello que no se puede pensar que no hay semántica en la estructura. Seguro que hay pistas de ella, como lo verifican algunos estudios a partir de representaciones de palabras (Arroyo-Fernández et al., 2018b), pero la idea transmitida codificada en un mensaje se resuelve mayoritariamente en el cerebro de los hablantes usando información precargada en este (ideas, imágenes, referentes).

"El gato ladra". En efecto, esta oración difícilmente crea una combinación con significado en la mayoría de los cerebros, aunque puede ser transmitida y es parte de un lenguaje. Además, puede ser fácilmente representada en un espacio vectorial usando cualquier método de embebido.

Otro ejemplo: en este tiempo existen formas de comunicación que asumen que los interlocutores tienen conocimiento previo del contexto de los mensajes. Por ejemplo, las aplicaciones de mensajería instantánea. En varias de estas, incluso existe un límite de caracteres que pueden ser enviados. Los interlocutores incluso se pueden comunicar sin que el mensaje esté completo o tenga errores ortográficos y gramaticales muy significativos.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Conclusiones

Se ha evaluado a nuestro método de representación de oraciones (WISSE) sobre un conjunto de tareas de STS bien conocidas. Nuestro método superó al estado del arte en varias de estas tareas. La más importante de ellas la constituye el conjunto de datos SICK, en la cual se alcanzó una correlación de $\rho = 0.724$. Esto es alentador porque este conjunto es relativamente difícil para los métodos no supervisados de representación de oraciones (de hecho, la barrera de $\rho = 0.72$ es difícil de superar).

WISSE superó también a los métodos del estado del arte en un par de conjuntos de datos (Answer-Answer y Plagio) del SemEval-2016. Estos conjuntos de datos constituyen tareas de STS especialmente complejas. Por ejemplo, para Answer-Answer nuestro método fue mejor en aproximadamente un 10 %. Para el resto de conjuntos de datos, el desempeño de WISSE se mantuvo en una vecindad del 2,7 % con respecto al estado del arte.

También comparamos WISSE contra una amplia variedad de métodos que participaron en el concurso SemEval-2016. Aunque muchos (la mayoría) de estos métodos son supervisados o incorporan conocimiento externo, WISSE superó a la media global. Los diagramas de caja mostraron que en la mayoría de las tareas de STS WISSE superó al 95 % de los competidores. Además, se mantuvo cerca (a menos de 4 %) del mejor sistema STS, el cual se basa en aprendizaje supervisado (regresión de vectores de soporte y aprendizaje profundo), así como en una gran variedad de recursos externos tales como WordNet, Wikipedia y reconocedores de entidades nombradas.

Nuestros experimentos confirmaron nuestra hipótesis, la cual afirma que es posible representar las frases usando el vínculo entre los contextos estimados por word embeddings y la entropía de las palabras que componen una oración. Hemos

explotado el enlace mencionado de manera que fue posible ajustar los coeficientes de una sumatoria de word embeddings para representar oraciones. De manera muy notable, tales coeficientes resultaron ser simples escalares. Este hecho, además de que permitió a nuestro modelo generalizar muy bien en tareas tareas STS de diversa naturaleza y a bajo costo computacional, también provee un panorama adicional acerca de la complejidad* que se espera debería caracterizar a una máquina de aprendizaje que modele el problema de STS. Esto no necesariamente implica que la tarea de STS es poco compleja por que sólo se requieran escalares para parametrizar a un modelo. De hecho, nuestra discusión sobre las propiedades del texto a representar muestra casos concretos de estas propiedaes para las cuales tanto WISSE como el estado del arte fallaron. Más bien, a partir de tal hecho se puede intuir que una tarea de STS promedio o de propósito muy general, no requiere una máquina de aprendizaje compleja.

La modularidad de nuestro modelo ofrece la posibilidad de configurarlo según las propiedades de texto, lo que nos permitió obtener el mejor desempeño en diversas tareas de STS. También es interesante el hecho de que tal desempeño se obtuvo mediante el uso de funciones de similitud simples tales como Coseno y Euclidiana, que están definidas en espacios euclidianos. Esto permite concluir que las presentaciones producidas por WISSE pueden considerarse ajustadas a este tipo de geometría. Desde luego, ello dependerá de las propiedades de la oración, como ya se vio en la discusión.

Finalmente, el bajo coste computacional y modularidad de WISSE resulta especialmente útil cuando se tiene solo texto sin etiquetar en aplicaciones de pocos datos y de inferencia de representaciones en tiempo real (online). Por ejemplo, representar y procesar los resultados de una consulta cualquiera en un motor de búsqueda.

5.2. Trabajo futuro, ventajas y desventajas

5.2.1. Fuentes ocultas de información

En los experimentos hechos hasta ahora sólo se consideró una versión básica de nuestro modelo usando un a estructura de tres niveles de información proporcionados por la entropía de Shannon (equivalente a TF–IDF). Los vectores IDF se obtuvieron de una forma ingenua que asume distribuciones uniformes de las oraciones, de las oraciones que comparten a cada elemento del vocabulario. Estos

*Esta complejidad puede ser cuantificada en términos de la *VC*-dimension, la cual es función del número de parámetros (o coeficientes) de la máquina y del número de patrones que se espera que esta debe aprender a separar con tal cantidad de parámetros (Vapnik, 1998).

supuestos podrían llevar al modelo propuesto a omitir cosas importantes que estén relativamente alejadas del promedio de los contextos. Entre las más importantes de ellas es la hipótesis explícita sobre una medida de probabilidad que subyace a los contextos y emisiones lingüísticas. Es claro que las oraciones que comparten una palabra dada no siguen distribuciones uniformes en el proceso estocástico de comunicación (como lo asume el estimador usado por WISSE).

Un experimento futuro interesante sería calcular un estimador de las densidades de probabilidad que subyacen la estructura de la información en un corpus. Sería interesante que este estimador considere más niveles de la estructura de información del corpus. Por ejemplo, es posible que un nivel adicional tome en cuenta, en cierta medida, las propiedades de un anillo algebraico (*ring*), como lo propuso Chomsky. Derbyshire (1977) propone que un hablante puede emitir la oración de la forma (1), de forma tal que se ha puesto énfasis en el agente ("Paula") de la acción ("golpeó"). Por el contrario, en la oración (2), el hablante ha puesto énfasis en el paciente ("Nacho"). Esta no es una regla general, pero seguro que de tener un estimado de ella sería posible tener una estructura de información con mucho mayor detalle sobre las cláusulas de una oración.

(1) *Paula golpeó a Nacho.*

(2) *Nacho fue golpeado por Paula.*

Entonces a nivel cláusula, se puede observar que la auto similitud de estos patrones de la lengua se jerarquiza y es recurrente en construcciones más complejas. Por ejemplo, en *"Paula golpeó Nacho porque su ofensa fue inadmisible"*, la frase nominal "su ofensa" en la cláusula subordinada es también enfatizada por el hablante; con respecto a lo "inadmisible" de la "ofensa". No obstante para la totalidad de la oración, todo énfasis que localmente toma lugar en la cláusula subordinada ("su ofensa") es jerárquicamente menos importante que cualquiera que pueda observarse en la cláusula principal ("Paula"). En general, se sabe que la clausula principal es más importante que la subordinada. Este tipo de patrones de información no necesariamente requieren un etiquetador sintáctico, sino más bien una técnica adecuada para segmentar los niveles de la jerarquía de conjuntos del corpus.

Por supuesto que lo expuesto en esta subsección dependerá de las exigencias de las aplicaciones del modelo. Los desempeños altos que se registraron indican que existen escenarios donde no se requiere mayor capacidad de representación (o más niveles en la estructura de la información).

5.2.2. Ventajas y desventajas.

Como lo puntualiza en la sección anterior, WISSE no codifica adecuadamente información específica. Este es el caso, por ejemplo, de los adverbiales de negación. En este sentido, las representaciones de las oraciones (3) y (4) serían prácticamente iguales:

(3) *Las computadoras no condenan a la humanidad.*

(4) *Las computadoras condenarán a la humanidad.*

Aunque se podría considerar que semánticamente estas dos oraciones son opuestas, estadísticamente, en la estructura de subconjuntos del corpus, la forma adverbial "no" se utiliza de la misma forma que una palabra funcional: con frecuencias similares y en contextos indiscriminados. Cada vez que el hablante tiene la necesidad de negar cualquier hecho. Así, desde el punto de vista de la estructura de información actualmente incorporada en WISSE, esta palabra es poco informativa. Dependiendo de la forma en que se calculen vectores TF (binarios, logarítmicos o frecuencias), los coeficientes de cualquier sumatoria que contenga a esta palabra no cambiarán gran cosa debido a su alta probabilidad de estar compartida por muchos contextos. Esta y otras "imprecisiones semánticas" son una desventaja considerable de WISSE actualmente, pero podrían abordarse mediante una estructura de subconjuntos más detallada. En la tabla 5.1 se muestran algunos coeficientes TF–IDF diferentes para la palabra *"not"* participando en un par de oraciones aleatoriamente extraídas de nuestro corpus STS.

Tabla 5.1: Coeficientes TF–IDF para la forma adverbial *"not"* participando de un par de oraciones.

Oración	TF-IDF
The man jumping is not wearing a shirt.	"jumping" (0.6537), "wearing" (0.4753), "shirt" (0.5679), "not" (**0.1894**)
A girl is close to a boy whose face is not shown.	"girl" (0.4258), "close" (0.3415), "boy" (0.4339), "face" (0.3833), "not" (**0.1959**)

Otra ventaja de la modularidad

En torno a esta discusión, se observa que la distinción de estructuras de subconjuntos en las muestras lingüísticas permite identificar varias categorías lingüísticas que están estadísticamente embebidas en WISSE. Esta identificabilidad es una ventaja significativa, ya que es posible vislumbrar la manipulación de la contribución de dichas categorías de acuerdo con la aplicación o tarea de NLP. Por

ejemplo, si se tuviera un interés por medir la similitud entre oraciones con respecto a la entidad "molécula proteica" (una frase nominal), basta con incrementar los coeficientes de la serie para dicha entidad a un valor adecuado (Badarinza et al., 2017). Esto permitiría medir *similitud semántica dirigida* a términos, por ejemplo.

Nótese que este tipo de identificabilidad es definitivamente muy complicada de alcanzar con otros métodos de representación en el estado del arte. La mayoría de ellos están basados exclusivamente en redes neuronales.

Anexos

Anexo A

Chomsky: "generativismo" o funciones generadoras

En diversas áreas dentro de la lingüística se puede observar un dogma sobre los trabajos de Noam Chomsky. Aunque esto no es general en la lingüística, sí es fácil observar puntos de vista poco amplios, que tanto defienden como rechazan a este autor. Este dogma surge a partir de la postura, no necesariamente filosófica pero tampoco probada, de que la lengua se genera a partir de una estructura discreta. Se toma esta afirmación general como aquella defendida por Chomsky como verdad única. Sin embargo, es posible que este dogma no esté del todo fundamentado. Primero porque este autor no defiende tal cosa. Y segundo porque cualquier científico define una teoría o representación basándose en un conjunto de principios, los cuales en muchos casos tienen la finalidad de facilitar un análisis en particular (es usual *delimitar un tema*). Por ejemplo, si se define una gramática en particular, no quiere decir dicha definición que esta gramática sea la que explica a todos los fenómenos del objeto de estudio. Lo mismo es válido para el caso en que se definan un conjunto de gramáticas. Estas tendrían por objeto facilitar el análisis de un conjunto de casos de interés; pero no todos los fenómenos o hipótesis posibles latentes en el objeto de estudio (las lenguas naturales, en este caso). Si se lee con atención a Chomsky and Miller (1958) o bien a Chomsky and Schützenberger (1963) es posible confirmar, al menos, que el dogma mencionado aquí carece de fundamento. En la sección C, se plantearán observaciones teóricas sobre la naturaleza estadística de las teorías de Chomsky.

En lingüística se suele decir también que las lenguas naturales son un fenómeno social exclusivamente. Si bien no es una postura general, es fácil entablar un debate al respecto en cualquier reunión con estudiosos del área. Lejos de tomar una u otra postura, el autor de esta tesis critica el hecho de que se tome una

u otra. Aquí la postura más antropocéntrica (la lengua solo como fenómeno social) ciertamente es bastante subjetiva. Se observa una tendencia generalizada a clasificar a Chomsky como representante de esta postura. Curiosamente, su trabajo fundamental no muestra en absoluto indicios que justifiquen tal *tradición*. Más bien, este transmite la idea de que en efecto la lengua puede ser un fenómeno social (principalmente originado a partir de las necesidades humanas primarias), pero que también, y principalmente, es un fenómeno que resulta de observar a un sistema físico como cualquier otro. En la misma línea, el facilitar un análisis en particular delimitando un tema mediante una contradicción a esto no quiere decir que tal contradicción en particular sea el trabajo de este autor. Generalmente se dice que Chomsky no cree en la estadística o en la incertidumbre, encasillando así a su trabajo como un tipo de análisis muy general de categorías discretas y reglas demasiado estrictas. Probablemente poco útil para la mayoría de los casos de una lengua. Esto sí es cierto si sólo se conocen o se piensa en las llamadas *gramáticas generativas*.

Es posible observar patrones estructurales en diversos fenómenos de nuestro entorno que no necesariamente se pueden observar mediante valores categóricos (como los símbolos con los que registramos usualmente a las lenguas naturales), sino más bien mediante valores en una escala continua (Elman, 1990, Hauser et al., 2002, Koulouris et al., 2006, Kowalski et al., 2012, Liu and Fu, 1983, Visnevski et al., 2007). Si se piensa en ello, también es fácil aceptar que la lengua originalmente no era escrita, o que un individuo no alfabetizado participa en un sistema de comunicación usando los mismos medios físicos que uno alfabetizado. Es precisamente este tipo de ideas de tinte estadístico las que parecen ser productos o *formalizaciones* de la influencia de su tutor, Z. B. Harris (Harris, 1954, 1957, 1968, 1991). En particular las funciones generadoras de estructura (*Structure Generating Functions*) introducidas por el mismo Chomsky materializan tal formalismo. Pero no como un formalismo de la lingüística formal, como generalmente se refiere a ello en las tradiciones, sino como un formalismo matemático y estadístico. No obstante, esto puede ser rebatible por la tradición, pues esta parece estar guiada solo por el recurrente estilo de conclusiones en los trabajos de Chomsky. Estas generalmente resultan de un método de análisis que primero busca patrones en las muestras lingüísticas y después los formaliza en un *conjunto de reglas*, mismas que no dejan de ser un *resumen* (como él mismo las llama) de dichos patrones. Nótese que un resumen es una versión recortada de los datos, una representación global, que seguramente omite los detalles de la descripción de los mismos (la teoría algebraica de los lenguajes libres de contexto) y que el autor en cuestión omitió cuando sus trabajos, la mayoría, estuvieron enfocados en aplicaciones (análisis lingüístico).

Sus ideas más detalladas y con mayor alcance descriptivo están en sus trabajos fundamentales. Más que como tradiciones *generativistas*, Chomsky transmite estas ideas como un modelo matemático de la generación de lenguajes, basado principalmente en la frecuencia y longitud de sus símbolos (y subsímbolos, recursivamente). Incluso puntualiza las similitudes de su trabajo con respecto al de Shannon. Estas se pueden apreciar al comparar los dos trabajos y sus resultados, cuyo nivel de representación y poder explicativo son de envergadura más que destacada (Chomsky and Miller, 1958, Chomsky and Schützenberger, 1963, Shannon, 1948). Son precisamente estas similitudes las que inspiraron muchos otros trabajos grandiosos que en la misma época revolucionaria tuvieron lugar (Banerji, 1963, Ginsburg, 1966, Kaminger, 1970, Kuich, 1970). Son estas similitudes comprobables también las que dan lugar a esta discusión, dado que la tradición lingüística incluso niega su existencia.

Existe mucho trabajo por hacer en investigación en lingüística computacional, NLP e inteligencia artificial, pero muchos huecos teóricos (actuales) serían menos pronunciados de tomar en cuenta varias investigaciones importantes que aún cuando no se hace mucho esfuerzo por entender, se consideran *"pasadas de moda"*.

Anexo B

Deep Learning y tendencias

Los métodos de representación de oraciones basados puramente en el aprendizaje profundo (*Deep Learning*) han mostrado resultados competitivos en tareas como clasificación de oraciones (por sentimiento u otros criterios) (Kalchbrenner et al., 2014). Sin embargo, para la tarea de medición de similitud semántica los métodos que tienen el desempeño más alto requieren supervisión (Conneau et al., 2017, Yin et al., 2016), lo que vuelve completamente prohibitivo su uso en aplicaciones para lenguas diferentes al inglés o para temáticas especializadas. Algunos de estos métodos tienen la limitación de que su costo computacional puede ser muy alto, además de que la cantidad de datos necesaria para entrenarlos es también muy alta, lo que limita mucho su uso en aplicaciones de propósito general (esto incluye a otros dominios de datos tales como imágenes).

Estos defectos no se deben necesariamente al lento avance de la tecnología de cómputo. Es decir, tener a disposición un gran equipo de cómputo de alto desempeño no significa que usarlo para analizar un problema cuya alta complejidad es cuestionable sea necesariamente la mejor solución. Esta observación se debe a que en este tiempo, en muchas áreas de la ciencia de datos, se eligen máquinas de aprendizaje extremadamente complejas en términos de su capacidad, en comparación con la capacidad que un problema de aprendizaje realmente demanda de una máquina (Vapnik, 1998). De manera indiscriminada se usan modelos con alta complejidad para analizar problemas de los que se sabe poco, lo que lleva muchas veces al sobreajuste (*overfitting*) o, en el mejor de los casos, a complejidades y poder de cómputo innecesarios para un modelo y/o problema de aprendizaje. Soslayar esta observación deja a la suerte el éxito en la generalización de una máquina sobre un problema, ya que no se sabe lo necesario sobre los datos del problema ni sobre la máquina a la que se le muestran.

Mucha de la investigación actual en aprendizaje automático (*Machine Learning*)

se basa en "modas" impuestas por algunas entidades preponderantes y de prestigio ciertamente dudoso (empresas de tecnología principalmente). Ello debilita el valor de hacer investigación sobre redes neuronales, ya que, como parte de la moda, los resultados de la nueva investigación se dejan a la suerte, más que a conocer los fundamentos teóricos que desde hace mucho se han estudiado sobre estos modelos (Cybenko, 1989, Hornik et al., 1989, Vapnik, 1998), pero que se quedan rápidamente bajo la sobra de la propaganda.

Por último, aunque estudiar los modelos no es de gran valor monetario, la historia es la evidencia cultural de que siempre habrá la necesidad de conocer y preservar las teorías así como de aportar nuevas, mismas que son inmunes a la caducidad. Las tecnologías, las modas y las preponderancias no gozan de tal inmunidad.

Anexo C

Indicios de estructura

En la figura C.1 se puede observar que la estructura de los subconjuntos de S también se manifiesta en estas fluctuaciones de entropía (figura C.2). En este ejemplo (un fragmento de "Alicia en el país de las maravillas"*), las palabras más informativas se alternan con las más entrópicas. Entre las más informativas, también hay otras que lo son menos y también se alternan. Lo mismo se observa con las palabras más entrópicas.

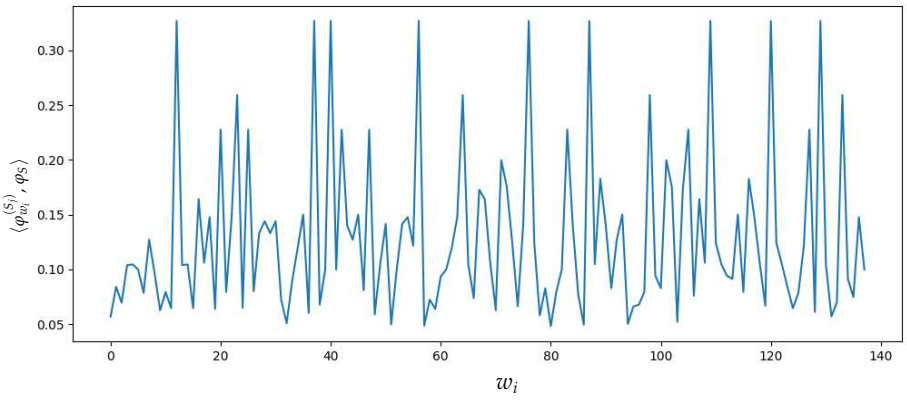

Fig. C.1: Fluctuaciones de entropía en un fragmento de "Alicia en el país de las maravillas".

Con base en las observaciones anteriores, se hicieron algunas inferencias que pretenden dar pistas sobre porqué la hipótesis distribucional ha favorecido un modelo tan empíricamente sencillo como el que se propone en esta tesis (y otros).

*Fragmento transformado: *"There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, Oh dear ! Oh dear! I shall be late! (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural); but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge."*

No se hará una demostración matemática por el momento, solo se citan hechos ya estudiados (algunos de ellos posiblemente ya han sido demostrados) en trabajos anteriores. En este sentido, trabajo de Z.S. Harris y de N. Chomsky es particularmente interesante. Esta discusión sobre indicios de estructura en la semántica distribucional, se basa en una proposición de Harris, misma que posteriormente fue formalizada por Chomsky:

"... some morphemes have very similar (though not identical) sets of co-occurrents: thus, the set of co-occurrents for 'cloth'– e.g. The ( ) tore, The ( ) was torn, Get me a ( ) quick– may have many morphemes in common with the set for 'paper', certainly many more than with the set for 'diminish'. This suggests that morphemes can be grouped into classes in such a way that members of a class have rather similar sets of co-occurrents, and each class in turn occurs with specific other classes to make a sentence structure.":
Z.S. Harris (Harris, 1957) (1957)

Se habla de una formalización desde el punto de vista matemático, no lingüístico (para ese entonces). Es decir, Chomsky, entre otras muchas cuestiones matemáticas, propone que la estructura se deriva (o deriva) patrones de coocurrencia individual. Esto lo define Harris:

*"The range of individual co-occurrence of a morpheme (or word) **i** is defined first of all as the environment of morphemes (or words) which occur in the same sentences with **i** (in some body of linguistic material).*
This is indeed the initial information available for morphological structure".
Z.S. Harris (Harris, 1957) (1957)

Aunque este trabajo de tesis no se concentra en fenómenos morfológicos, en realidad Chomsky generaliza estas observaciones en un modelo matemático formal de estructura que no requiere especificar si se desea estudiar algún nivel de análisis de la lengua en particular (fonología, morfología o sintaxis). Dicho modelo matemático estudia la estructura de secuencias de símbolos que asu vez constituyen cadenas r, que pueden producir *cadenas terminales* $f_i \in V_t$, donde V_t es un vocabulario. Por ejemplo, tanto la coocurrencia individual de palabras como la de los conjuntos de ellas puede verse como una serie de potencias ponderadas, lo que a su vez constituye un *proceso generativo* de la forma (C.1):

$$r = \sum_i \langle r, f_i \rangle f_i. \tag{C.1}$$

Las cadenas terminales (e.g. las palabras) constituyen a su vez un lenguaje

L, para el cual r es una *representación estructural*[*] de la secuencia $f_1, f_2, \ldots, f_i, \ldots$. Chomsky establece que este proceso es generado por una gramática G, luego los f_i se dice que tienen asociados *descriptores estructurales* derivados de G (e.g. árboles sintácticos que básicamente aglomeran elementos, cadenas terminales, en categorías jerárquicas definidas por las subcadenas que sus elementos tienen en común). De esta manera, se dice que L es generado por el proceso generativo descrito por G, i.e. $L(G)$. Por lo tanto, si algún $\langle r, f_i \rangle = 0$, entonces la correspondiente cadena terminal f_i no puede ser descrita por G (o no está en el *soporte de* G o bien $f_i \notin L(G)$). Por el contrario, si $\langle r, f_i \rangle > 0$ entonces se confirma que $f_i \in L(G)$, ya que f_i puede ser descrita mediante G.

El coeficiente de cada cadena terminal de L es muy importante, independientemente del tipo de estructura que describa G (un proceso generativo está latente en muchos fenómenos naturales muy interesantes (Collado-Vides, 1992, Howell et al., 1980, Liu and Fu, 1983, Visnevski et al., 2007), además de las lenguas naturales). De manera más específica, si se define al número de descriptores estructurales de f como $N(G, f) = \langle r(G), f_i \rangle$ y mediante el análisis de f se determina que $N(G, f) > 1$, entonces f puede ser descrito por más de un descriptor estructural. El hecho de que más de un descriptor estructural (p. ej., 2 o 3 árboles sintácticos), esté asociado a una misma oración f implica *ambigüedad estructural*. $N(G, f) = 1$ implica, entonces, no ambigüedad. Por ahora se dejan de lado los casos en que $N(G, f) < 0$ (lo que define las gramáticas dependientes del contexto, no aplicables a las lenguas naturales).

Por un lado, es fácil observar que la ecuación (3.7) que se obtuvo como modelo de oración en esta tesis es bastante similar a la *función generadora de estructura* (C.1). Se llama ahora así porque ya se definió el fenómeno que modela, por lo que ya no es solo una serie de potencias que modela a un proceso generativo. Por otro lado, no es muy fácil observar que ambos modelos pueden ser conectados mediante una analogía entre entropía y ambigüedad estructural. Es mucho menos fácil aún probar matemáticamente que dicha analogía no es tal, sino más bien una relación en concreto (una tarea aparentemente ardua que se dejará como trabajo futuro). Se argumenta dicha analogía en términos de un ejemplo sencillo de gramáticas, que al mismo tiempo coincide con la hipótesis de coocurrencia y estructura de Harris que se mencionó al principio. Véase cómo. Sea una gramática y su correspondiente

[*]No confundir la acepción que se usa justo aquí de la palabra "representación" con la acepción que se ha usado a lo largo de la tesis en cuanto vectores o embebidos de palabras u oraciones. La primera es una representación matemática, mientras que la segunda es una representación numérica (en forma de vector) del significado.

ecuación

$$\begin{aligned} S \rightarrow\ & SbS; S \rightarrow a \\ S =\ & a + SbS. \end{aligned} \tag{C.2}$$

Aquí la gramática describe a un proceso generativo general modelado por su ecuación. Es posible mostrar cómo este proceso generativo continua hasta obtener un subconjunto del lenguaje que es capaz de generar. Para ello, se define la función $\phi(r) = a + rbr$. Nótese que r ya está definida en (C.1). Con ello en mente, el proceso generativo continúa de manera recursiva, generando nuevos niveles jerárquicos que ahora se denotan como ϱ_i:

$$\begin{aligned} \varrho_0 =& r_0 = 0 \\ \varrho_1 =& r_1 = a + r_0 b r_0 = a + 0b0 = a \\ \varrho_2 =& r_2 = a + r_1 b r_1 = a + aba \\ \varrho_3 =& r_3 = a + r_2 b r_2 = a + (a + aba)b(a + aba) \\ &= a + (a + aba)(ba + baba) \\ &= a + aba + ababa + ababa + abababa \\ &= a + aba + 2ababa + abababa \\ \varrho_4 =& r_4 = a + r_3 b r_3 \\ &= a + aba + a(ab)^2a + 5(ab)^3a + 6(ab)^4a + 6(ab)^5a + 4(ab)^6a + (ab)^7a. \\ &\ddots \end{aligned} \tag{C.3}$$

Los polinomios ϱ_i cumplen con las propiedades matemáticas de un anillo topológico (*ring*). Este anillo es un conjunto de elementos (números o símbolos) que cumplen con todas las propiedades algebraicas de un espacio vectorial métrico, excepto por el hecho de que *el producto no es conmutativo*. Como en cualquier entorno algebraico, los exponentes representan repeticiones del mismo patrón, e.g. $(ab)^3 = ababab$; asimismo, estas expresiones algebraicas se pueden descomponer en factores, como en el caso de ϱ_3.

En esta sección se ha puesto especial atención en el hecho curioso de que el proceso generativo descrito por G sea a su vez un sistema de polinomios en el que las potencias de estos sean repeticiones de cadenas terminales. Otro hecho curioso es que si se observa de manera general a la ecuación (C.3), de arriba hacia abajo, se ve cómo los elementos del conjunto formado por ϱ_0 usan a r_i para co-ocurrir con el contexto formado por las cadenas terminales de ϱ_1. Esto se reproduce para los siguientes conjuntos más grandes $i = 2, 3, \ldots$. Con base en ello, se llega al punto

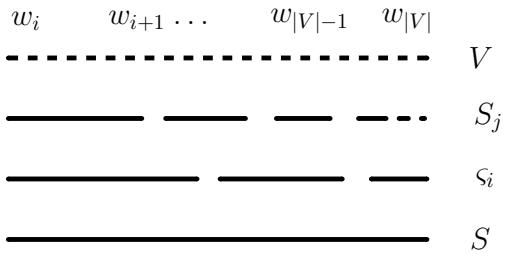

Fig. C.2: Estructura (*bottom-top*) de subconjuntos de palabras identificados en el corpus de texto.

principal de esta discusión. La subcadena $r_1 b = ab \in \varrho_2$ no coocurre en la primera cadena terminal a de ϱ_4; pero, coocurre una vez en la segunda cadena terminal aba de ϱ_4; coocurre dos veces en la tercera cadena terminal $a(ab)^2a$; coocurre tres veces en la cuarta cadena terminal $a(ab)^3a$, etc. Esto, explicado a detalle en el trabajo de Chomsky, confirma la hipótesis de Harris que relaciona directamente a la estructura y a la coocurrencia (mencionada al inicio de esta sección).

Bien, se ponen a disposición las preguntas que resultan de esta interpretación del trabajo de Chomsky-Harris en el sentido de procesos generativos*. ¿En términos de coocurrencia (repetición), es posible vislumbrar ahora que también este fenómeno puede ser observado en una matriz término-documento? ¿Es la frecuencia (o repetición) de las cadenas terminales que produce un modelo generativo lo que tenemos en dicha matriz? Entonces, ¿*Latent Semantic Analysis* no solo es una representación cuantitativa sino también estructural del lenguaje modelado? Desde el punto de vista de las funciones generadoras de estructura de Chomsky, TF–IDF es de hecho una transformación definida sobre los exponentes de una gramática bastante grande. Tales exponentes, al ser el número de repeticiones de una cadena en un lenguaje, pueden ser utilizados para medir jerarquías de probabilidades (como se mostró en el caso de la ecuación (3.4)).

Los coeficientes de la suma ponderada (3.7) provienen de las frecuencias de coocurrencia de las cadenas terminales que comparten los diferentes conjuntos de oraciones de la estructura de 4 niveles $w_i \in S_j \subseteq \varsigma_i \subseteq S$, como en la figura C.2. Estas a su vez se transforman en probabilidades. Estas probabilidades a su vez indican qué tan informativas o entrópicas son las cadenas terminales. Por lo tanto, en términos de un proceso generativo de estructura, el exponente entero de cada monomio de un polinomio se ha transformado en la probabilidad de dicho monomio y luego en su entropía. Nótese entonces que la relación entre estas cosas constituye una transformación sobre los exponentes del polinomio que modela

*"Proceso generativo" es un término muy utilizado en este tiempo y con la misma acepción en la literatura de *Machine Learning*.

dicho proceso:

$$\begin{aligned} \varrho(x_j; j) =& a_0 + a_1 x_1 + a_2 x_2^2 + \cdots + a_i x_j^j + \cdots + a_n x^n \\ &\mapsto P(x_j) = \frac{j}{\sum_{j=1}^{n} j} \\ &\mapsto H(x_j) = -\sum_j P(x_j) \log P(x_j). \end{aligned} \tag{C.4}$$

Entonces, a partir de (C.4) nos preguntamos si los coeficientes de la serie ponderada por entropía (3.7), en el dominio de $H(x)$, representan un criterio alternativo de ambigüedad en comparación con el definido para $N(G, f)$ en el dominio de los coeficientes de $\varrho(x_j; j)$. Un acercamiento a la respuesta de esta pregunta se plantea por analogía. Según Chomsky, mientras mayor sea el número $N(G, f)$ de descriptores estructurales derivados para una cadena terminal f a partir de la gramática G, más ambigüedad estructural se tendrá. Asimismo, desde el punto de vista de Shannon, mientras más contextos comparta una cadena terminal x_j con otras x ($x_j \in \varrho_i \forall i$), más entrópica será esta en términos de su significado. Por lo tanto, observando la transitividad de (C.4), teóricamente ambos tipos de estructura (entropía y gramática) están consideradas en los coeficientes de la serie ponderada por entropía (3.7) a través de una transformación $F \mapsto H$. Esta va desde el dominio del proceso generador de estructura de las cadenas terminales al dominio del proceso generador de estructura de su entropía:

$$F(\varrho)(j) = H(x_j) = \quad -\sum_j \frac{j}{\sum_{j=1}^{n} j} \log \frac{j}{\sum_{j=1}^{n} j}. \tag{C.5}$$

Puesto que se ha definido la relación entre los conjuntos $\varrho(x)$ y $P(x)$, todas las operaciones $H(x)$ definidas en la sección 2.4 sobre los conjuntos de tipo $P(x)$ tienen sentido estructural.

Algunas de estas preguntas aquí planteadas pueden ser difíciles de responder empíricamente, puesto que en realidad los principios introducidos por Chomsky están delimitados, por él mismo, a lenguajes artificiales y son más útiles para esos casos (por ejemplo, un lenguaje de computadora debe cumplir con $N(G, f) = 1, \forall f \in V_t$, esto es, no debería ser estructuralmente ambiguo*).

Una última consideración sobre los componentes de nuestro modelo trata con los word embeddings. Desde el punto de vista de la ecuación (C.4), los word embeddings están el dominio de la probabilidad $P(x_j)$, pero cumplen con el mismo esquema. Esto porque, como ya se mencionó antes, cada término de cada polino-

*Por supuesto, esto es importante porque si no se prueba para las ordenes que recibe un sistema inteligente, las consecuencias podrían ser fatales si este tiene a su cargo vidas humanas.

mio $\varrho(x; j)$ puede ser a su vez (recursivamente) producido por un conjunto de funciones $\phi(r) = a + rbr$; es decir, otros polinomios que coocurren en $\varrho_1, \ldots, \varrho_i, \ldots,$. Para el caso de los word embeddings, los exponente de tales polinomios provienen del conteo de intersecciones recursivas entre las ventanas de contexto unas contra otras. Dichos conteos son transformados en probabilidades de coocurrencia. Aunque tales ventanas de contexto en la mayoría de los casos no cumplen estrictamente con los requerimientos de una gramática, estadísticamente sí forman estructuras contextuales, o jerarquías de coocurrencia (Allenby and Rossi, 2006, Huang and Madan, 1999, Hyndman et al., 2011), que permiten distinguirlas como muestras o segmentos generados por procesos generadores de estructura. Estos conteos inducen estimadores que pueden ser usados para calcular grados de dependencia e independencia estadística entre las palabras a partir de sus contextos (similitud en semántica distribucional). Se puede entonces intuir que la longitud de las ventanas de contexto es un parámetro relacionado con los niveles jerárquicos del proceso generativo de estructura que las describe. En este sentido, un lenguaje no sólo se construye con palabras, sino también con segmentos de ellas, $n-$gramas de segmentos de ellas, $n-$gramas de ellas y así sucesivamente. Cualquier conjunto de cadenas (y subcadenas) terminales. Como la estructura no necesariamente obedece reglas (estas solo son resúmenes de un comportamiento más complejo), los procesos generativos de estructura están latentes en los datos textuales solo por la distribución de sus elementos (fonemas, morfemas, palabras, oraciones, discursos, etc.), misma que está concebida principalmente para transportar información*.

*¿Qué tipo de lenguaje resultaría si sus cadenas terminales tuvieran distribución uniforme? ¿y con la biología?

Anexo D

Otro modelo propuesto

Durante el desarrollo de este trabajo ha sido posible visualizar de manera general varios algoritmos de embebido para diversas fuentes de generación de datos, además de las lenguas humanas. Se identificaron diversos (meta)patrones que comparten varios algoritmos. Los más sobresalientes de estos patrones metodológicos fueron la idea del aprendizaje autodidacta, el cual puede ser extrapolado a cualquier generador de datos para modelar su comportamiento, y la estructura jerárquica de entropías en TF–IDF. En este sentido, y si reescribimos el modelo general propuesto en este trabajo en términos de distribuciones de probabilidad, se proponen un par objetivos de entrenamiento autodidacta para estimar representaciones de oraciones desde cero. El primer modelo adicional propuesto tiene la forma:

$$\mathcal{J}(x_w, x_{c_k}) = \sum_{\substack{c_k \in D \\ w \in V}} \sum_{k=1}^{K} \sum_{w \in c_k} P(w|c_k; x_{c_k}, x_w) \log \frac{P(c_k|w; x_{c_k}, x_w)}{P(c_k)} \log P(w|c_k; x_{c_k}, x_w). \tag{D.1}$$

El segundo tiene la forma:

$$\mathcal{J}(x_w, x_{c_k}) = \sum_{\substack{c_k \in D \\ w \in V}} P(w|c_k; x_{c_k}, x_w) \log \frac{P(c_k|w; x_{c_k}, x_w)}{P(c_k)}. \tag{D.2}$$

Los parámetros x_w, cuando el costo $\mathcal{J}(x_w, x_{c_k})$ es óptimo, pueden ser utilizados para inferir representaciones de contextos definidas de dos maneras: la más sencilla es calcular simplemente

$$x_{c_k} = x_{w_1} +, \cdots, + x_{w_{|c_k|}} = \sum_{w_i \in c_k} x_{w_i}.$$

Otro método, aunque mucho más costoso (pero tal vez mejor), consiste en calcular la máxima probabilidad posterior $P(c_k|w; x_{c_k}, x_w)$:

$$x_{c_k} = \arg_{x_c} \max_{\alpha_c} \sum_{w \in c_k} \log P(c|w; x_c, x_w) = \arg_{x_c} \max_{\alpha_c} \sum_{w \in c_k} \log \frac{1}{Z} e^{x_w^\intercal x_c}$$

donde

$$x_{c_k} = \alpha_c^\intercal C_k = \sum_{w \in c_k} \alpha_w x_w, \tag{D.3}$$

se conoce como *operador se síntesis*, $C_k \in \mathbb{R}^{K \times d}$ es la matriz de representaciones de K palabras de un contexto c_k y $\alpha_c \in \Lambda \subset \mathbb{R}^K$ es un vector de coeficientes estimado a partir del contexto c_k, cuyas componentes ponderan a las representaciones de palabras correspondientes. Cada α_c pertenece a un conjunto de hipótesis Λ, donde el óptimo es α_{c_k}. Este puede ser estimado por una RNN que recorre las K representaciones la matriz de contexto C_k:

$$z^{(i)}(x_{w_i}) = Y_f^\intercal f_A^{(i-1)} + Y_c^\intercal x_{w_i} + b \tag{D.4}$$

donde

$$f_A^{(i-1)}(x_{w_i}) = \sigma[z^{(i-1)}(x_{w_{i-1}})],$$

$Y_f \in \mathbb{R}^{K \times K}$ es la matriz de coeficientes de atención, $Y_c \in \mathbb{R}^{d \times K}$ es la matriz de análisis de representaciones, $\sigma(\cdot)$ es la distribución de probabilidad $P(c|w; x_c, x_w)$, y una vez que la red está entrenada $\alpha_{c_k}(C_k) = f_A^{(K)} \in \mathbb{R}^K$ para $i = 1, \ldots, K$. Esta red se considera entrenada cuando la función de costo (D.2) es óptima, lo que mantiene sin supervisión la inferencia de α_c para contextos no vistos. Se ha mostrado en trabajos previos que una RNN no es opción cuando se busca embeber representaciones de contenido semántico (Arroyo-Fernández and Meza Ruiz, 2017), pero en este caso se desea que la red aprenda a predecir atención en las representaciones de los contextos. Esto implica un modelo de *reconstrucción*, la *síntesis* de x_{c_k} (D.3) a partir del *análisis* de contenido de los átomos x_w (Elad et al., 2007). Este enfoque que distingue síntesis de análisis requiere mucho menos complejidad (computacional y del modelo) que uno de estimación de contenido semántico que no distinga entre estas dos operaciones. Aquí, (D.4) incluye la fase de análisis que solo usa las x_w como entradas para predecir su jerarquía en el contexto, pero no pretende embeberlas nuevamente (lo cual sí sería una tarea muy compleja para este tipo de red neuronal). Nótese que las probabilidades posteriores se han modelado con la distribución de Bolzmann (también llamada *Softmax*, en este tiempo).

Estos métodos de construcción de contextos también puede ser usados como

métodos para construir representaciones de oraciones. Usando la suma simple: $x_{S_j} = x_{w_i} +, \cdots, + x_{w_{|S_j|}}$. Y calculando la máxima probabilidad posterior:

$$x_{S_j} = \sum_{w \in S_j} \log P(S_j | x_j, x_w) = \sum_{w \in S_j} \log \frac{1}{Z} e^{\alpha_j^\intercal C_j},$$

donde la constante de normalización Z se define generalmente como

$$Z = \sum_{c \in D} e^{\alpha_j^\intercal C_j}.$$

En (D.2) se ha suprimido el logaritmo de la verosimilitud de la distribución de palabras dados los contextos $\log P(w|c_k; x_{c_k}, x_w)$. Esto porque una hipótesis plausible indica que la forma de la Información mutua que se modela, la divergencia de Kullback-Leibler ponderada por $w \in V$, entre las distribuciones de palabras y de contextos satisface la estructura de representación requerida. Esto es, se requiere representar oraciones dado que se observan cada una de las palabras que contienen (la heurística detrás de TF–IDF, enfocada en factores discriminadores entre contextos). No al revés. Esto se interpreta a partir de la asimetría de (D.2), donde las distribuciones a priori $P(w|c; x_c, x_w)$ y $P(w)$ ponderan a la distribución $P(c|w; x_c, x_w)$ (la interpretación de Aizawa (2003)). Se omite, sin embargo, el logaritmo de la verosimilitud de la distribución de palabras dados los contextos, que puede ser un factor crítico ya que esta se concentra en las representaciones de las palabras que forman a la oración (la serie ponderada propuesta en este trabajo).

Una interpretación adicional ayuda a confirmar la viabilidad del modelo. Se trata de la operación de convolución entre las distribuciones de probabilidad. Aquí, la distribución de los contextos dado el vocabulario funciona como proyector (un filtro estocástico); mientras que la distribución del vocabulario es la señal filtrada por el proyector (Principe, 2010). Con ello también adquiere lógica el hecho de que se asuman distribuciones de Boltzmann, ya que pueden interpretarse como filtros pasa bajas cuyos parámetros son las representaciones (frecuencia de corte, velocidad de respuesta, etc.). Algo también bastante interesante de esta convolución es que opera sobre objetos puramente categóricos y no ordenados, escogiéndolos; mientras que las convoluciones tradicionales lo hacen necesariamente sobre objetos en conjuntos ordenados y de manera determinista (enteros, reales, complejos, etc.). Este punto de vista ha sido poco explorado (Erdogmus et al., 2005), pero generalmente sobre conjuntos ordenados discretizados; con fines de simulación.

Nótese que los objetivos (D.1) y (D.2) entrenan simultáneamente representaciones de palabras y de contextos, por lo que será interesante verificar varios niveles

de representación. Por lo pronto, algo muy importante de remarcar es que, al nivel de palabras, estos objetivos entrenan representaciones de palabras para satisfacer la necesidad específica de información en la tarea de representación de oraciones. La optimización de estos objetivos se puede llevar a cabo mediante métodos de estimación, como la Maximización de la verosimilitud y de la expectación. Otros métodos, son la estimación Bayesiana y los multiplicadores de Lagrange. En general con estos se busca resolver numéricamente un sistema de la forma:

$$\begin{aligned} \nabla_{Y_f, Y_c} \mathcal{J}(\alpha_c, x_w) &= 0 \\ \nabla_{x_w} \mathcal{J}(x_c, x_w) &= 0 \end{aligned} \tag{D.5}$$

Este sistema de ecuaciones mide los cambios de verosimilitud para dos clasificadores, que comparten datos pero que los ven desde puntos de vista antagónicos: el primero, $P(c|w; x_c, x_w)$, es un clasificador que calcula la probabilidad posterior de un c que discrimina a un conjunto determinado de palabras. Nósese que este clasificador está parametrizado por α_c y no por Y_f y Y_c (Eq. (D.4)). En este caso, el gradiente mide la magnitud del vector de Información mutua para un conjunto disperso de coeficientes α_c, estimados mediante la RNN de parámetros Y_f y Y_c. Estos últimos pueden ser actualizados usando una regla delta que penaliza el valor esperado del retorno de entropía de las representaciones x_{S_j}, dadas las estimaciones (Sutton and Barto, 2018, Sutton et al., 2000):

$$Y_{(\cdot)}^{(t)} \leftarrow Y_{(\cdot)}^{(t-1)} - \beta^{(t-1)} \nabla_{Y_{(\cdot)}} \mathbb{E}[\mathcal{J}(\alpha_c, x_w)],$$

donde $\beta^{(t)} \in \mathbb{R}$ es una taza de aprendizaje (*learning rate*) que puede variar en cada iteración t. El segundo clasificador, $P(w|c; x_c, x_w)$, es el que usualmente se tiene en métodos de representaciones distribuidas de palabras (CBoW). Este discrimina los contextos en que aparece una palabra y está parametrizado por x_w. Cada clasificador está parametrizado por un conjunto de parámetros por separado, pero ambos clasificadores están entrelazados por la función de costo*. Desde un punto de vista estadístico, los dos clasificadores optimizan los parámetros de dos distribuciones cuya intersección es la más informativa entre los contextos. Para esto se plantea el sistema de ecuaciones (D.5) con costo asimétrico (D.2). Por ahora, en efecto, se deja como trabajo futuro la implementación y revisión detallada de este modelado.

Algunas restricciones para el costo D.2 pueden ser incorporadas a la estimación de parámetros tomando en consideración las observaciones de De Marcken (1999). Este autor propone un modelo bastante elegante de gramáticas a partir de

*Aunque comparten parámetros, no se tienen clasificadores siameses. Solo modelan distribuciones de eventos con parámetros complementarios.

diferencias de entropía en la oración. Esto, por un lado resuelve el problema de la supresión del logaritmo de la verosimilitud, al menos hipotéticamente. Por otro lado, esto ya incorporaría un nivel más detallado y generalizado de representación que ya no se limitaría únicamente a temáticas. Claro, incluir tal capacidad de representación al modelo deprenderá de las necesidades de información de cada aplicación.

Cabe señalar que estos modelos propuestos se basan en la hipótesis de que consideran una estructura de conjuntos muy similar a la del modelo general (3.1), i.e. $w_i \in c_k \subseteq \zeta_i \subseteq D$. Esto les permitiría representar objetos y composiciones de estos en cualquier entorno, dado que no se limitan a datos de texto (admiten cualquier conjunto de elementos contables que se desee representar en un espacio de embebido).

Bibliografía

N. Afzal, Y. Wang, and H. Liu. Mayonlp at semeval-2016 task 1: Semantic textual similarity based on lexical semantic net and deep learning semantic model. *SemEval NAACL-HLT*, pages 674–679, 2016.

E. Agirre, M. Diab, D. Cer, and A. Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393, 2012.

E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo. * sem 2013 shared task: Semantic textual similarity. *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, 1:32–43, 2013.

E. Agirre, C. Banea, D. Cer, M. Diab, A. Gonzalez-Agirre, R. Mihalcea, G. Rigau, and J. Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. *Proceedings of SemEval*, pages 497–511, 2016.

A. Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

R. Alarcón, G. Sierra, and C. Bach. Developing a definitional knowledge extraction system. *Conference Proceedings of Third Language & Technology Conference LTC'07*, 2007.

G.M. Allenby and P.E. Rossi. Hierarchical bayes models. *The handbook of marketing research: Uses, misuses, and future advances*, pages 418–440, 2006.

G. Angeli, M.J.J. Premkumar, and C.D. Manning. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint*

Conference on Natural Language Processing (Volume 1: Long Papers), volume 1, pages 344–354, 2015.

S. Arora, Y. Liang, and T. Ma. A simple but tough-to-beat baseline for sentence embeddings. *International Conference on Learning Representations (ICLR)*, 2017.

I. Arroyo-Fernández. Agrupamiento semántico de definiciones en un espacio generado mediante aprendizaje de kernels. Phd research project, Universidad Nacional Autónoma de México, 2013.

I. Arroyo-Fernández. The Describe Corpus: A recopilation of text snippets containing sense definitions retrieved from the web and their emebddings, 2016. URL `http://github.com/iarroyof/describe_corpus`. Grupo de ingeniría lingüística – Universidad Nacional Autónoma de México.

I. Arroyo-Fernández and I.V. Meza Ruiz. LIPN-IIMAS at semeval-2017 task 1: Subword embeddings, attention recurrent neural networks and cross word alignment for semantic textual similarity. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 199–203, August 2017.

I. Arroyo-Fernández, C.F. Méndez-Cruz, G. Sierra, J.M. Torres-Moreno, and G. Sidorov. Unsupervised sentence representations as word information series: Revisiting tf–idf. *arXiv preprint arXiv:1710.06524*, 2017.

I. Arroyo-Fernández, D. Forest, J.M. Torres-Moreno, M. Carrasco-Ruiz, T. Legeleux, and K. Joannette. Cyberbullying detection task: the EBSI-LIA-UNAM system (ELU) at COLING'18 TRAC-1. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 140–149, 2018a.

I. Arroyo-Fernández, I. Meza, and C.F. Méndez-Cruz. UNAM at semeval-2018 task 10: Unsupervised semantic discriminative attribute identification in neural word embedding cones. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 977–984, 2018b.

I. Arroyo-Fernández, J.M. Torres-Moreno, G. Sierra, and L.A. Cabrera-Diego. Automatic text summarization by non-topic relevance estimation. *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management: KDIR*, 1:89–100, 2016.

I. Badarinza, A.I. Sterca, and M. Ionescu. Syntactic indexes for text retrieval. *INFORMATION TECHNOLOGY IN INDUSTRY*, 5:24–28, 2017.

R. Balan, P.G. Casazza, C. Heil, and Z. Landau. Density, overcompleteness, and localization of frames. i. theory. *Journal of Fourier Analysis and Applications*, 12(2): 105–143, 2006.

R.B. Banerji. Phrase structure languages, finite machines, and channel capacity. *Information and Control*, 6(2):153–162, 1963.

M. Baroni and A. Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *ACL (1)*, pages 238–247, 2014.

Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

L. Bentivogli, R. Bernardi, M. Marelli, S. Menini, M. Baroni, and R. Zamparelli. Sick through the semeval glasses. lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Language Resources and Evaluation*, 50(1):95–124, 2016.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

C.D. Boom, S.V. Canneyt, S. Bohez, T. Demeester, and B. Dhoedt. Learning semantic similarity for very short texts. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1229–1234, 2015.

E. Bruni, N. Tram, M. Baroni, et al. Multimodal distributional semantics. *The Journal of Artificial Intelligence Research*, 49:1–47, 2014.

T. Brychcın and L. Svoboda. Uwb at semeval-2016 task 1: Semantic textual similarity using lexical, syntactic, and semantic information. *Proceedings of SemEval*, pages 588–594, 2016.

H. Calsamiglia and A. Tusón. *Las cosas del decir: manual de análisis del discurso*. Ariel, 1999.

H. Calsamiglia Blancafort and A. Tusón Valls. *Las cosas del decir: Manual de análisis del discurso*. Ariel, Barcelona, first edition, 1999.

D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, August 2017.

E. Charniak. *Statistical language learning*. MIT press, 1996.

H. Chen, S.S. Fuller, C. Friedman, and W. Hersh. Knowledge management, data mining, and text mining in medical informatics. In *Medical Informatics*, pages 3–33. Springer, 2005.

T. Chen, R. Xu, Y. He, and X. Wang. Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Systems with Applications*, 72: 221–230, 2017.

K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

N. Chomsky and G.A. Miller. Finite state languages. *Information and Control*, 1(2): 91 – 112, 1958.

N. Chomsky and M.P. Schützenberger. The algebraic theory of context-free languages. *Studies in Logic and the Foundations of Mathematics*, 35:118–161, 1963.

J. Collado-Vides. Grammatical model of the regulation of gene expression. *Proceedings of the National Academy of Sciences*, 89(20):9405–9409, 1992.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

L.M. Davidson. *Knowledge extraction technology for terminology.* PhD thesis, University of Ottawa (Canada), 1998.

C. De Boom, S. Van Canneyt, T. Demeester, and B. Dhoedt. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156, 2016.

C. De Marcken. On the unsupervised induction of phrase-structure grammars. *Natural Language Processing Using Very Large Corpora*, pages 191–208, 1999.

D.C. Derbyshire. Word order universals and the existence of ovs languages. *Linguistic Inquiry*, 8(3):590–599, 1977.

P. Dhillon, D.P. Foster, and L.H. Ungar. Multi-view learning of word embeddings via cca. In *Advances in neural information processing systems*, pages 199–207, 2011.

P.S. Dhillon, D.P. Foster, and L.H. Ungar. Eigenwords: spectral word embeddings. *Journal of Machine Learning Research*, 16:3035–3078, 2015.

R.J. Duffin and A.C. Schaeffer. A class of nonharmonic fourier series. *Transactions of the American Mathematical Society*, 72(2):341–366, 1952.

M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse problems*, 23(3):947, 2007.

J.L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

J.L. Elman. Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, 7(2-3):195–225, 1991.

M.J. Er, Y. Zhang, N. Wang, and M. Pratama. Attention pooling-based convolutional neural network for sentence modelling. *Information Sciences*, 373:388–403, 2016.

D. Erdogmus, R. Agrawal, and J.C. Principe. A mutual information extension to the matched filter. *Signal Processing*, 85(5):927 – 935, 2005. Information Theoretic Signal Processing.

A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, 2011.

R.M. Fano. *The transmission of information*. Massachusetts Institute of Technology, Research Laboratory of Electronics Cambridge, Mass, USA, 1949.

J. Ferrero, L. Besacier, D. Schwab, and F. Agnès. Compilig at semeval-2017 task 1: Cross-language plagiarism detection methods for semantic textual similarity.

Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 100–105, August 2017.

J.R. Firth. *Papers in linguistics 1934-1951*. Oxford University Press, London, 1957.

J. Fourier. *Theorie analytique de la chaleur*. Chez Firmin Didot, père et fils, 1822.

S.L. Frank. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, 5(3):475–494, 2013.

K. Fukushima. Neural network model for selective attention in visual pattern recognition and associative recall. *Applied Optics*, 26(23):4985–4992, 1987.

S. Ginsburg. *The Mathematical Theory of Context Free Languages.[Mit Fig.]*. McGraw-Hill Book Company, 1966.

G.H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische mathematik*, 14(5):403–420, 1970.

I.J. Goodfellow, D. Warde-Farley, M. Mirza, A.C. Courville, and Y. Bengio. Maxout networks. *ICML (3)*, 28:1319–1327, 2013.

J. Hale. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123, Mar 2003.

J. Hale. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672, 2006.

Z.S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Z.S. Harris. Co-occurrence and transformation in linguistic structure. *Language*, 33(3):283–340, 1957.

Z.S. Harris. *Mathematical Structures of Language*. Wiley, New York, NY, USA, 1968.

Z.S. Harris. *Theory of Language and Information: A Mathematical Approach*. Oxford University Press UK, 1991.

R.V. Hartley. Transmission of information. *Bell Labs Technical Journal*, 7(3):535–563, 1928.

V. Hatzivassiloglou, J.L. Klavans, and E. Eskin. Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. *Proceedings of the 1999 joint sigdat conference on empirical methods in natural language processing and very large corpora*, pages 203–212, 1999.

M.D. Hauser, N. Chomsky, and W.T. Fitch. The faculty of language: What is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579, 2002.

C. Heil. History and evolution of the density theorem for gabor frames. *Journal of Fourier Analysis and Applications*, 13(2):113–166, 2007.

F. Hill, K. Cho, and A. Korhonen. Learning distributed representations of sentences from unlabelled data. *Proceedings of NAACL-HLT*, pages 1367–1377, 2016.

G. Hinton, J. McClelland, and D. Rumelhart. Distributed representations. *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*, pages 77–109, 1986.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

J. Howell, T. Smith, and M. Waterman. Computation of generating functions for biological molecules. *SIAM Journal on Applied Mathematics*, 39(1):119–133, 1980.

X. Huang and A. Madan. Cap3: A dna sequence assembly program. *Genome research*, 9(9):868–877, 1999.

D.A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

R.J. Hyndman, R.A. Ahmed, G. Athanasopoulos, and H.L. Shang. Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, 55(9):2579–2589, 2011.

I. Iacobacci, M.T. Pilehvar, and R. Navigli. Sensembed: Learning sense embeddings for word and relational similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 95–105, 2015.

W. Jarosz, N.A. Carr, and H.W. Jensen. Importance sampling spherical harmonics. *Computer Graphics Forum*, 28(2):577–586, 2009.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 487–494. Omnipress, 2010a.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 366–373, 2010b.

Y. Ji and J. Eisenstein. Discriminative improvements to distributional sentence similarity. *EMNLP*, pages 891–896, 2013.

D. Jurafsky and J.H. Martin. *Speech and language processing*, volume 3. Pearson, London, 2014.

D. Jurgens, M.T. Pilehvar, and R. Navigli. Cross level semantic similarity: an evaluation framework for universal measures of similarity. *Language Resources and Evaluation*, 50(1):5–33, 2016.

N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, June 2014.

F. Kaminger. The noncomputability of the channel capacity of context-sensitive languages. *Information and Control*, 17(2):175–182, 1970.

T. Kenter and M. de Rijke. Short text similarity with word embeddings. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1411–1420, 2015.

J.D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9. Association for Computational Linguistics, 2009.

M. King, W. Gharbieh, S. Park, and P. Cook. UNBNLP at semeval-2016 task 1: Semantic textual similarity: A unified framework for semantic processing and evaluation. *Proceedings of SemEval*, pages 732–735, 2016.

W. Kintsch and P. Mangalath. The construction of meaning. *Topics in Cognitive Science*, 3(2):346–370, 2011.

R. Kiros, Y. Zhu, R.R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. *Advances in neural information processing systems*, pages 3294–3302, 2015.

A. Koulouris, T. Andronikos, C. Pavlatos, A. Dimopoulos, I. Panagopoulos, and G. Papakonstantinou. Efficient signal processing using syntactic pattern recog-

nition methods. In *International Conference on SIGNAL AND IMAGE PROCESSING, Honolulu, Hawaii, USA*, 2006.

A.M. Kowalski, M.T. Martin, A. Plastino, and G. Judge. On extracting probability distribution information from time series. *Entropy*, 14(10):1829–1841, 2012.

W. Kuich. On the entropy of context-free languages. *Information and Control*, 16(2): 173–200, 1970.

S. Kullback. *Information theory and statistics*. Courier Corporation, 1997.

T.K. Landauer, P.W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284, 1998.

Q. Le and T. Mikolov. Distributed representations of sentences and documents. *31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1188–1196, 2014.

D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

O. Levy and Y. Goldberg. Dependency-based word embeddings. *ACL (2)*, pages 302–308, 2014.

H.H. Liu and K.S. Fu. An application of syntactic pattern recognition to seismic discrimination. *IEEE Transactions on Geoscience and Remote Sensing*, GE-21(2): 125–132, 1983.

T.M. MacRobert. *Spherical harmonics. An elementary treatise on harmonic functions with applications*. Third edition revised with the assistance of I. N. Sneddon. International Series of Monographs in Pure and Applied Mathematics, Vol. 98. Pergamon Press, Oxford, 1967.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696. ACM, 2009.

B. Mandelbrot. *Les objets fractals : forme, hasard et dimension, survol du langage fractal*. Flammarion, 4th edition edition, 1999.

C.D. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval*. Cambridge UP, 2009.

C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

D.I. Martin and M.W. Berry. Mathematical foundations behind latent semantic analysis. *Handbook of latent semantic analysis*, pages 35–56, 2007.

T. Maszczyk and W. Duch. Comparison of shannon, renyi and tsallis entropy used in decision trees. In L. Rutkowski, R. Tadeusiewicz, L.A. Zadeh, and J.M. Zurada, editors, *Artificial Intelligence and Soft Computing – ICAISC 2008*, pages 643–651, Berlin, Heidelberg, 2008.

I. Meyer. Extracting knowledge-rich contexts for terminography. *Recent advances in computational terminology*, 2:279, 2001.

I. Meza-Ruiz and S. Riedel. Jointly identifying predicates, arguments and senses using markov logic. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 155–163. Association for Computational Linguistics, 2009.

R. Mihalcea, C. Corley, C. Strapparava, et al. Corpus-based and knowledge-based measures of text semantic similarity. *AAAI*, 6:775–780, 2006.

T. Mikolov, I. Sutskever, A. Deoras, H.S. Le, S. Kombrink, and J. Cernocky. Subword language modeling with neural networks. *preprint (http://www. fit. vutbr. cz/imikolov/rnnlm/char. pdf)*, 2012.

T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

G.A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

J. Mitchell and M. Lapata. Composition in distributional models of semantics. *Cognitive Science*, 34(34):1388–1429, 2010. Cognitive Science Society, ISSN: 1551-6709.

J. Mueller and A. Thyagarajan. Siamese recurrent architectures for learning sentence similarity. *AAAI*, pages 2786–2792, 2016.

P. Neculoiu, M. Versteegh, M. Rotaru, and T.B. Amsterdam. Learning text similarity with siamese recurrent networks. *ACL 2016*, page 148, 2016.

M.E. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

A. Onan, S. Korukoğlu, and H. Bulut. A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing & Management*, 53(4):814–833, 2017.

D.B. Osteyee and I.J. Good. *Expected mutual information*, pages 26–38. Springer Berlin Heidelberg, Berlin, Heidelberg, 1974.

S. Padó and M. Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

M. Pagliardini, P. Gupta, and M. Jaggi. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. *arXiv*, 2017.

T. Pedersen, A. Purandare, and A. Kulkarni. Name discrimination by clustering similar contexts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 226–237. Springer, 2005.

J. Pennington, R. Socher, and C.D. Manning. Glove: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

F. Pereira. Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 358(1769):1239–1253, 2000.

N.T. Pham, G. Kruszewski, A. Lazaridou, and M. Baroni. Jointly optimizing word representations for lexical and sentential tasks with the c-phrase model. *ACL (1)*, pages 971–981, 2015.

M.T. Pilehvar and R. Navigli. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95 – 128, 2015.

J.C. Principe. *Information theoretic learning: Rényi's entropy and kernel perspectives*. Springer, 2010.

R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM, 2007.

P. Rastogi, B. Van Durme, and R. Arora. Multiview lsa: Representation learning via generalized cca. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–566, 2015.

A. Rényi. On measures of entropy and information. Technical report, HUNGARIAN ACADEMY OF SCIENCES Budapest Hungary, 1961.

S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520, 2004.

X. Rong. word2vec parameter learning explained. *CoRR*, abs/1411.2738, 2014.

S. Rothe and H. Schütze. Word embedding calculus in meaningful ultradense subspaces. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 512–517, 2016.

D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McClelland, and C. PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

B. Rychalska, K. Pakulska, K. Chodorowska, W. Walczak, and P. Andruszkiewicz. Samsung poland nlp team at semeval-2016 task 1: Necessity for diversity; combining recursive autoencoders, wordnet and ensemble methods to measure semantic similarity. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, CA, USA*, pages 602–608, 2016.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

G. Salton, E.A. Fox, and H. Wu. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036, 1983.

M. Sato, H. Ogawa, and T. Iijima. A theory of pseudo-orthogonal bases and its application to image transmission. In *Applications of Digital Image Processing VI*, volume 432, pages 38–45. International Society for Optics and Photonics, 1984.

B. Schölkopf, P. Simard, A. Smola, and V. Vapnik. Prior knowledge in support vector kernels. *Proceedings of the 10th International Conference on Neural Information Processing Systems*, pages 640–646, 1997.

C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 & 623–656, 1948.

C.E. Shannon. Communication theory of secrecy systems*. *Bell system technical journal*, 28(4):656–715, 1949.

C.E. Shannon. A symbolic analysis of relay and switching circuits. Master's thesis, Massachusetts Institute of Technology, 1940.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge UP, 2004. ISBN: 978-0-521-81397-6.

G. Sierra. Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *LinguaMÁTICA*, 2:13–38, Dezembro 2009.

K. Spärk Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

M. Sugiyama and M. Kawanabe. *Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation*. Adaptive computation and machine learning. MIT Press, 2012.

M. Sugiyama and H. Ogawa. Pseudo orthogonal bases give the optimal generalization capability in neural network learning. In *Wavelet Applications in Signal and Image Processing VII*, volume 3813, pages 526–538. International Society for Optics and Photonics, 1999.

M.A. Sultan, S. Bethard, and T. Sumner. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*, 2:219–230, 2014.

R.S. Sutton and A.G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

R.S. Sutton, D.A. McAllester, S.P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

M. Těšitělová. *Quantitative linguistics*. Linguistics and literary studies in Eastern Europe. Academia Publishing House of the Czechoslovak Academy of Sciences, 1992.

R. Tian, N. Okazaki, and K. Inui. The mechanism of additive composition. *Machine Learning*, 106(7):1083–1130, Jul 2017.

C. Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52(1):479–487, Jul 1988.

V.N. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton. Grammar as a foreign language. *Advances in Neural Information Processing Systems*, pages 2773–2781, 2015.

N. Visnevski, V. Krishnamurthy, A. Wang, and S. Haykin. Syntactic modeling and signal processing of multifunction radars: a stochastic context-free grammar approach. *Proceedings of the IEEE*, 95(5):1000–1025, 2007.

J.H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII*, pages 499–515. Springer International Publishing, Cham, 2016.

P. Wiemer-Hastings. All parts are not created equal: Siam-lsa. In *Proceedings of the Cognitive Science Society*, volume 27, 2005.

J. Wieting, M. Bansal, K. Gimpel, and K. Livescu. Charagram: Embedding words and sentences via character n-grams. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1515, November 2016.

T. Williamson. Sense, validity and context. *Philosophy and Phenomenological Research*, 57(3):649–654, 1997.

Y. Xu, M.Y. Kim, K. Quinn, R. Goebel, and D. Barbosa. Open information extraction with tree kernels. *HLT-NAACL*, pages 868–877, 2013.

M. Yazdani and A. Popescu-Belis. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. *Artificial Intelligence*, 194: 176–202, 2013.

W. Yin and H. Schütze. Discriminative phrase embedding for paraphrase identification. *Proceedings of HLT-NAACL*, pages 1368–1373, 2015.

W. Yin, H. Schütze, B. Xiang, and B. Zhou. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272, 2016.

J. Yu, L. Xie, X. Xiao, and E.S. Chng. Learning distributed sentence representations for story segmentation. *Signal Processing*, 2017.

Z. Zhang, S.S. Ge, and H. He. Mutual-reinforcement document summarization using embedded graph based sentence clustering for storytelling. *Information Processing & Management*, 48(4):767–778, 2012.

G. Zheng and J. Callan. Learning to reweight terms with distributed representations. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 575–584, 2015.