



UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS
INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS
CENTRO DE INVESTIGACIÓN EN CIENCIAS

**Estudio de un modelo predictivo de relaciones asimétricas
aplicado a la representación de rasgos lingüísticos**

T E S I S
QUE PARA OBTENER EL TÍTULO DE
LICENCIADO EN CIENCIAS
(COMPUTACIÓN)

PRESENTA:
ELISEO MORALES GONZALEZ

DIRECTOR DE TESIS
DR. JORGE HERMOSILLO VALADEZ

CUERNAVACA, MORELOS

13 de junio de 2021



UNIVERSIDAD AUTÓNOMA DEL
ESTADO DE MORELOS



Instituto de
Investigación en
Ciencias
Básicas y
Aplicadas

INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS



Control Escolar de Licenciatura

VOTOS DE APROBATORIOS

Secretaria ejecutiva del Instituto de Investigación en Ciencias Básicas Aplicadas de la Universidad Autónoma del Estado de Morelos.

P r e s e n t e .

Por medio de la presente le informamos que después de revisar la versión escrita de la tesis que realizó el C. **ELISEO MORALES GONZÁLEZ** con número de matrícula **20164007732** cuyo título es:

"Estudio de un modelo predictivo de relaciones asimétricas aplicado a la representación de rasgos lingüísticos"

Consideramos que **SI** reúne los méritos que son necesarios para continuar los trámites para obtener el título de **LICENCIADO EN CIENCIAS ÁREA TERMINAL EN CIENCIAS COMPUTACIONALES Y COMPUTACIÓN CIENTÍFICA.**

Cuernavaca, Mor a 04 de junio del 2021

Atentamente
Por una universidad culta

Se adiciona página con la e-firma UAEM de los siguientes:

DR. MARKUS MÜLLER BENDER
DRA. LORENA DÍAZ GONZÁLEZ
DR. JORGE HERMOSILLO VALADEZ
DRA. ASELA REIG ALAMILLO
DR. GERARDO MAURICIO TOLEDO ACOSTA

PRESIDENTE
SECRETARIO
VOCAL
PRIMER SUPLENTE
SEGUNDO SUPLENTE



UNIVERSIDAD AUTÓNOMA DEL
ESTADO DE MORELOS

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

LORENA DIAZ GONZALEZ | Fecha:2021-06-04 12:31:23 | Firmante

LfR/Mhq33X6JaoJreRLsIRwkVz3UNzPoLEikj7fnJ7LLEq8qW1o6V9rF9Ahl+Z0tAY/UzPhQZDQWKA45L5Nn3s1TZ0dx5Rw2J0HRVTIqjBy3B30WwHNu+yxq/6swuyLc4Zx65BSf oHUDN3LNwobtlNzPy+eKkz/6yMIDaP+fzlxkoSdRGHxfH5/hjpZ0qM2JNwbw62PFMzha/4e0ZmnmBYu2yGT10ds1t1p9khYHhk2Pac4TrNYwgb27ZF8V7ScmUHV+MYrZEgNlSks 62Mfa53Bx1tRgQh0lai7kz4pEycmf/3H/u4dWkvHQY4A3N3QegDwDYiixdVgO1UuXEVOhw==

MARKUS FRANZISKUS MULLER | Fecha:2021-06-04 12:37:16 | Firmante

B5Xlea1xpULJfv13i7+0PJPII62Ws924Vt/4cCFwpCc2/B3AulLb5D0qw1fJcUTyY7qxohQJDBavjnv7Rf3sisuqVxDm+5+eeShkW1jY9HRUCxUttfZWz3n4HNmKrX3Q7XUZjHjRx+0o md2Sp2RsXVpLKl+Oe3qVgtn4FCjvLxaoCP9xuiFZsFys25ftUe1HDn0Ch5s9XmSOgJ+JFvpry9j+JEio+62VcukNUbZhzjRqHAYLwV+e0qzmlEjgvgINAASLw4YMPyEA7XkdtX3fi L5rFIMHEdl+rW4PyTLVi9E0BxjPcAQAI59AX2gNajCSgNoVbj0xGde9UmqixhA==

GERARDO MAURICIO TOLEDO ACOSTA | Fecha:2021-06-04 12:46:35 | Firmante

aW1QDuk5Yyu1M1pWhjDrGNdtOmznVd71qKDqKqGpUqzeaUCSfryd3Mnieum74co9RQtixD/ELM6kulTT8YW81Ski6KyAq9zulHOHNEjq/Z6hEdt3z3b6BMLSXiUjEabmE2GjXF W+rZBa/qSPAamAHKjGllnwhpZDdZ28C0AozPi/38/cR0GISTXmw2yTXSRh3RclPqngMeWXXdHRWRgkYrUXegB0saivJUMISXhTeTVJOhcvg+awXg6PMed1Uz9urX1EggHfwQ TonhT28yUwKUXgVgJ6ekFoNqEOHd20NukKZZ7Mk9AzfHwaNa5T/c3l8lfqD4OGU/zH8+1yw==

JORGE HERMOSILLO VALADEZ | Fecha:2021-06-04 16:36:43 | Firmante

I5MYV0i85lohi9QTCpGbnz+YUWkpDwBgSYp/8YgGx8DEdUN91PpmCWXQOvRtEoHvzcD5rG4ZBw/LQ5dlCuSL5h6HyJctt1BJ4ttmH9dlfmGixrFRwNviXN1rrFhoaL6oo7nipbrBn QFX5y+s5iMXPKiQmjypQbt913dVU15CoAgGdVtPMM/fNT3yC1MWWQLw3blkGlogOYTtERPereEiyLQMoh6HF5cCBNnIYAyrhnLkSh5Tf3u4+8xC5o2O25xU5esWtKXEFKynWk9 bLFrA6+cMysBvpjXC5AsE2yUcjTbChfGHOHHzCxt3nErB1uzri9+sTVkRr/barbFk/P201w==

MARIA ASELA REIG ALAMILLO | Fecha:2021-06-04 18:31:30 | Firmante

fIZ+/CsK8LADmUd3fwuJB+uY5FbohR965B9WY3kNAFX05D9OX0RrYNZeKxc0YAwMgj/K80emKnkASP3c2DnjCSEInecEf1eKiG5nokLpfcwN47qz51z0hCLXMLM+3HdYBVCd M7NGHGdwQSY05frTNXx2W0GEFH+37Ky/tymP0dt2oNVVU0/ovVThl5dNYM+BBuF61BAeR5Wcdr2UQ1y01N5b23epUfwpUt43YeLVmUKOMrK41+ol9Jk4xX4CSgGJ8e/kkPcY fVhLDJ6dLcwbDjdz3ZXn0h4Oetl7CzICt8fBpPulLymUKITkVWGSf+5losdnbhzhEk6lysJYreQ==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:



4CfAUD

<https://efirma.uaem.mx/noRepudio/18gmqRfaPfMkb5KPhFxiOicL60SLomJk>



Agradecimientos

En estas líneas quiero agradecer a todas las personas que hicieron posible esta investigación y que de alguna manera estuvieron conmigo en los momentos difíciles, alegres, y tristes. Estas palabras son para ustedes.

- A mi director de tesis, a quien le debo gran parte de mi aprendizaje y mi gusto por la investigación en el área, y junto a mis compañeros.
- A mi familia por todo su amor, comprensión y apoyo, pero sobre todo gracias infinitas por la paciencia que me han tenido. No tengo palabras para agradecerles las incontables veces que me brindaron su apoyo en todas las decisiones que he tomado a lo largo de mi vida, unas buenas, otras malas.

RESUMEN

Estudio de un modelo predictivo de relaciones asimétricas aplicado a la representación de rasgos lingüísticos

ELISEO MORALES GONZALEZ

La construcción de representaciones vectoriales de textos no es una tarea fácil. Existen muchos modelos que crean estas representaciones y cada modelo extrae cierto tipo de rasgos latentes involucrados en el texto. La mayoría de los modelos sólo se basan en la coocurrencia de palabras y no exploran otros rasgos lingüísticos que un texto posee. En este proyecto de tesis se realizó un estudio de las relaciones discursivas cuyo propósito fue observar las propiedades que poseen bajo un enfoque computacional, aplicando representaciones de texto que extraen los rasgos lingüísticos con un grado de abstracción mayor que el derivado de coocurrencias léxicas. Se presentan dos estudios: en el primero se realiza el modelo de las relaciones discursivas utilizando sus rasgos léxicos (categorías gramaticales) y se analizan sus propiedades con la ayuda de un modelo que construye representaciones vectoriales de palabras o texto; en el segundo estudio, se analizan las relaciones discursivas utilizando sus rasgos semánticos en términos de la sinonimia y antonimia existente entre sus argumentos. Este estudio se realizó sobre un corpus de relaciones discursivas llamado PDTB-3, específicamente, se trabajó con relaciones de categoría contraste. Los resultados nos indican que existe una diferencia entre las relaciones discursivas, y esta diferencia es dada por el tipo y el marcador discursivo que actúa como conector entre sus argumentos. En el segundo estudio, se observó que la mayoría de las relaciones discursivas carecen de sinonimia-antonimia entre sus argumentos. Sin embargo, se observó una considerable cantidad de relaciones discursivas que sí poseen este rasgo semántico. También se observó que existe una diferencia, no tan marcada, entre relaciones discursivas con marcadores implícitos y explícitos, las relaciones implícitas presentan una mayor presencia de sinonimia-antonimia que las relaciones explícitas.

Índice general

CAPÍTULO

1. Introducción	1
1.1. Antecedentes	1
1.2. Planteamiento del problema	4
1.2.1. Modelado de relaciones	5
1.3. Estado del arte	6
1.4. Pregunta de investigación	9
1.5. Hipótesis	9
1.6. Objetivos	9
1.6.1. Objetivo general	9
1.6.2. Objetivos específicos	9
2. Marco teórico	11
2.1. Modelo de estudio: ComplEx	11
2.2. Red semántica WordNet	13
2.3. Técnica de reducción de dimensionalidad: t-SNE	13
3. Metodología	15
3.1. Corpus de Relaciones discursivas	15
3.2. Señales léxicas	17
3.3. Representación de argumentos	18
3.4. Relación semántica: sinonimia-antonimia	19
3.5. Conjunto de entrenamiento para cada señal léxica	19
3.6. Propiedades geométricas de las incrustaciones	20
4. Resultados	23
4.1. Relaciones discursivas	23
4.2. Visualización de las propiedades geométricas	24
4.3. Señales léxicas	24
4.4. Señales semánticas: sinonimia-antonimia	26
5. Conclusiones	29
Bibliografía	31

Índice de figuras

1.1.	Árbol de dependencias gramaticales	4
3.1.	Ejemplo de los niveles (separados por un punto) para la categoría Comparación . .	16
3.2.	Ejemplos de conjuntos por señales léxicas.	17
3.3.	Representaciones por señal.	20
3.4.	Ejemplos de las propiedades geométricas.	21
4.1.	Resumen de tipo de relación por marcador discursivo y señal léxica	23
4.2.	Gráficas de violín utilizando proyecciones, descritas en la Sección 3.6.	25
4.3.	Incrustaciones de entidades para el marcador “on the other hand” de las señales verbo y adverbio.	26
4.4.	Incrustaciones de entidades para el marcador “on the other hand” (señales sustantivo y adjetivo).	26
4.5.	Incrustaciones de marcadores discursivos.	27
4.6.	Gráfica de los puntos (n_{ant}, n_{sin}) por marcador discursivo. Mientras más fuerte sea el color más puntos caen en la misma posición.	27
4.7.	Mapa de calor del conteo de puntos (n_{ant}, n_{sin}) separando marcadores explícitos e implícitos. Cada columna está marcada por tipo explícito (E) o implícito (I). . . .	28

Índice de tablas

1.1. Ejemplo del cálculo TF-IDF correspondiente a 4 términos presentes en una colección de 806 documentos, el cálculo TF-IDF se realiza para los documentos Doc1, Doc2 y Doc3.	3
1.2. Visualización de construcción de conjunto de entrenamiento del modelo CBOW	4
1.3. Ejemplos de relaciones gramaticales entre palabras	4
1.4. Ejemplos de tripletas MD para el sustantivo knife	6
3.1. Conjunto de relaciones discursivas	16

1.1. Antecedentes

En el Procesamiento de Lenguaje Natural (PLN), uno de los problemas a resolver es el de la representación de texto. El problema del significado y su modelación computacional, tiene sus raíces históricas en la Filosofía y la Inteligencia Artificial. En el PLN han surgido varios enfoques para modelar el lenguaje natural. El primer enfoque que surgió es el enfoque racional el cual busca explicar el funcionamiento del lenguaje, basándose en reglas o bases de conocimiento. El principal representante de este enfoque es Avram Noam Chomsky quien postula que la habilidad lingüística es innata y que sus bases sintáctico-semánticas están conectadas al cerebro al nacer como parte de la herencia genética humana. Sin embargo, aprender algo tan complejo como el lenguaje es difícil debido a que el lenguaje natural es ambiguo y su significado depende de muchos factores. Debido a que el lenguaje cambia y evoluciona con el tiempo surge la hipótesis de que existen patrones de uso de lenguaje y que estos patrones nos permiten interpretar o dar significado a las palabras, de esta manera surge el enfoque empirista. Este enfoque supone que el cerebro puede asociar, reconocer patrones y, por lo tanto, se pueden aprender las estructuras del lenguaje, de esta manera surgen los modelos basados en el enfoque estadístico del PLN, en el cual, mediante el Aprendizaje Automático es posible aprender las estructuras del lenguaje a partir de un corpus de textos.

Bajo el enfoque estadístico, existen dos formas de modelar el lenguaje natural. A la primera se le conoce como el Modelo de Lenguaje, mientras que a la segunda se le conoce como Modelos Distribuidos. El modelo de lenguaje consiste en la predicción de una palabra dada una secuencia de palabras previamente observadas. A este tipo de modelo se le conoce también como Modelo de N-gramas [1], matemáticamente se expresa como la descomposición de la distribución conjunta de N palabras siguiendo la regla del producto de la Teoría de la Probabilidad:

$$P(w_1, w_2, \dots, w_N) = P(w_1)P(w_2|w_1) \cdots P(w_N|w_1, w_2, \dots, w_{N-1}).$$

Cada elemento de la descomposición se le conoce como N-grama. Concretamente, a $p(w_1)$ se le llama unigrama, a $p(w_2|w_1)$ un bigrama, a $p(w_3|w_2, w_1)$ un trigramas, etc. La hipótesis de este modelo es que la primera palabra de un texto es independiente de cualquier otra, la segunda palabra depende de la primera palabra, la tercera de las dos primeras palabras y así sucesivamente; la última palabra depende de todas las anteriores.

En aplicaciones prácticas, este modelo se implementa usando sólo unigramas o bigramas. A la implementación usando solamente unigramas se le conoce como Modelo de Bolsa de Palabras (Bag of Words o BOW, por sus siglas en inglés). La implementación usando solamente bigramas consiste en definir una ventana de dos palabras que se va recorriendo a lo largo del texto para ir calculando las probabilidades de cada bigrama, a estas ventanas se les conoce como ventanas de contexto. De esta forma, es posible capturar la dependencia entre palabras en contextos de dos palabras. El tamaño

pequeño de las ventanas de contexto se debe a que las probabilidades se calculan usualmente por conteos de palabras, de tal suerte que el valor de probabilidad por cada tipo de N-grama disminuye drásticamente a medida que el tamaño del N-grama aumenta, ya que cada vez resulta más difícil encontrar ventanas mayores a 3 palabras cuyo contenido sea exactamente el mismo a lo largo de todo un corpus de texto. Es por esta razón que el modelo BOW es casi omnipresente.

Los modelos de lenguaje usando ventanas de tamaño menores a 3 se han aplicado de manera abundante en la literatura del PLN o Lingüística Computacional [2]. Sin embargo, estos modelos tienen dos serias limitaciones que son de gran relevancia para el presente trabajo. La primera es que los contextos que capturan estos modelos son usualmente tan pequeños como para capturar las relaciones entre palabras, las cuales se manifiestan entre palabras que se encuentran muy lejanas en la oración. La segunda y quizá más fuerte limitante es que estos modelos no permiten modelar relaciones semánticas entre palabras. Este es el interés de los modelos distribuidos.

Los modelos distribuidos de palabras se basan en la hipótesis de la semántica distribucional: “palabras con significado semejante, co-ocurren en contextos semejantes” (Harris, 1954, p.156).

En lugar de representar a las palabras como una distribución de probabilidad, estos modelos buscan generar vectores densos (vectores no nulos y con entradas reales) de dimensión N . La idea es contar con una estructura de espacio vectorial, en donde es posible realizar operaciones algebraicas y geométricas, que permiten dar cuenta de la relación semántica entre palabras, en función de cómo se asocian en el texto, mediante la cuantificación de la cercanía de sus vectores correspondientes.

Una forma clásica de hacer un modelo distribuido es mediante la representación TF-IDF [3], del inglés Term Frequency – Inverse Document Frequency, la cual es una medida numérica que expresa cuán relevante es un término ó palabra para un documento en un corpus de textos. La representación TF-IDF está dado por:

$$tfidf(t, d) = idf(t, D) * tf(t, d),$$

donde t es la palabra, d el documento y D es la colección de documentos. La expresión $idf(t, D)$ es la frecuencia inversa de documentos D , el cual nos dice que tanto el término t es frecuente en la colección de documento D . La expresión $tf(t, d)$ es la frecuencia de término t .

Por ejemplo, para modelar 3 documentos, se construye una matriz término-documento, donde cada fila es un término y cada columna un documento (ver Tabla 1.1 centro). De esta forma, se calculan los valores TF e IDF para luego aplicar la siguiente fórmula en cada casilla de la matriz (Ver Tabla 1.1 última).

Una ventaja de este tipo de modelos es que también se logran representaciones vectoriales de palabras, tomando en cuenta las filas. Otra ventaja, es que es posible definir una métrica sobre estos vectores, como la distancia Euclídeana; al definir una métrica se puede cuantificar la semejanza de documentos mediante la cercanía de sus vector. Las desventajas de los modelos TF-IDF y por Bolsa de Palabras son:

- La dimensión de los vectores es muy grande.
- La semejanza entre documentos o palabras es relativamente baja en casos reales, ya que depende esencialmente de que los documentos semejantes contengan las mismas palabras en una cantidad suficiente como para que los vectores estén cerca entre sí.

Cálculo IDF		
<i>Término</i>	<i>N</i>	<i>IDF</i>
carro	806	2.65
archivo		3.08
documento		2.50
museo		2.62

Frecuencia de término <i>t</i> en documento <i>i</i> TF (<i>i</i> =1,2,3)			
<i>Término</i>	<i>Doc1</i>	<i>Doc2</i>	<i>Doc3</i>
carro	27	4	24
archivo	3	33	0
documento	14	0	17
museo	0	33	29

Cálculo de pesos TF-IDF			
carro	tfidf(carro,Doc1) $27*2.65 = 71.55$	tfidf(carro,Doc2) $4*2.65 = 10.60$	tfidf(carro,Doc3) $24*2.65 = 63.60$
archivo	tfidf(archivo,Doc1) $3*3.08 = 9.24$	tfidf(archivo,Doc2) $33*3.08 = 101.64$	tfidf(archivo,Doc3) $0*3.08 = 0$
documento	tfidf(documento,Doc1) $14*2.50 =$	tfidf(documento,Doc2) $0*2.50 = 0$	tfidf(documento,Doc3) $17*2.50 = 42.50$
museo	tdidf(museo, Doc1) $0*2.62 = 0$	tfidf(museo, Doc2) $33*2.62 = 86.46$	tdidf(museo, Doc3) $29*2.62 = 75.98$

Tabla 1.1: Ejemplo del cálculo TF-IDF correspondiente a 4 términos presentes en una colección de 806 documentos, el cálculo TF-IDF se realiza para los documentos Doc1, Doc2 y Doc3.

Ante estos problemas, surgen los modelos de representación latente como Latent Semantic Analysis [4], que logran generar vectores densos en dimensiones relativamente bajas y que agrupan vectores de palabras semejantes, estableciendo semejanzas mucho más claras.

En los últimos años, han surgido modelos basados en redes neuronales que permiten lograr este tipo de representaciones de palabras, mediante vectores densos. Mediante la aplicación de Redes Neuronales Artificiales en el PLN se han propuesto modelos para crear representaciones incrustadas de palabras. Tal es el caso de Word2Vec [5] el cual tiene dos implementaciones, cada una con ciertas ventajas y desventajas:

- El modelo Continuo de Bolsa de Palabras (CBOW por sus siglas en inglés) que predice la palabra actual dado el contexto. Esta implementación se desempeña mejor para analogías sintácticas.
- El modelo Skip-Gram que predice el contexto circundante dada la palabra actual. Esta implementación funciona mejor en analogías semánticas.

Si se desea representar aspectos más finos del lenguaje, el modelo Word2Vec está limitado como se ilustra en la Tabla 1.2. En esta tabla se muestra un ejemplo de dos frases, una de las cuáles

Texto entrada	Palabra objetivo	Ventana contexto, $k = 2$
Ayer mi tío anduvo en bicicleta	anduvo	tío, bicicleta
Ayer mi bicicleta anduvo en tío	anduvo	bicicleta, tío

Tabla 1.2: Visualización de construcción de conjunto de entrenamiento del modelo CBOW

carece de sentido pragmático; sin embargo, para Word2Vec el modelado sería el mismo para cada una de las palabras, ya que el contexto es idéntico en ambas construcciones sintácticas.

Este ejemplo ilustra el problema que se estará tratando en este trabajo, el cual se describe en detalle en la siguiente sección.

1.2. Planteamiento del problema

En textos escritos en lenguaje natural, las palabras están relacionadas entre si ya sea mediante relaciones sintácticas, relaciones gramaticales o relaciones semánticas. La Figura 1.1 nos muestra el árbol de dependencias gramaticales [6] y la estructura de la frase “en la película viernes 13, el personaje Jason usó un cuchillo afilado para...”. Por ejemplo, la tripleta (*cuchillo*, *amod*, *afilado*) extraída de la Figura 1.1 nos indica que “cuchillo” se relaciona con “afilado” mediante la relación *amod*, es decir, que “afilado” es un adjetivo que modifica al sustantivo “cuchillo”. La Tabla 1.3 muestra la extracción de 2 tripletas, en la primera tripleta la relación involucrada es *amod* que significa “adjetivo modificador” y en la segunda, la relación involucrada es *obj* que significa “objeto de verbo”.

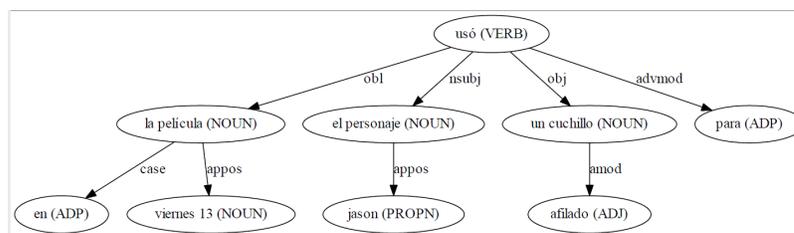


Figura 1.1: Árbol de dependencias gramaticales

Sujeto	Relación	Objeto
cuchillo	amod	afilado
cortar	obj	cuchillo

Tabla 1.3: Ejemplos de relaciones gramaticales entre palabras

El ejemplo de la Tabla 1.2 muestra claramente que existe una relación entre las palabras que, dependiendo de la estructura sintáctica de la frase, una de ellas tendrá más sentido que la otra desde un punto de vista semántico. Por ejemplo, la relación entre “tío” y “bicicleta” tendrá sentido si “tío” es sujeto de “anduvo” y “bicicleta” es objeto de “anduvo”, pero carecería de sentido si

“bicicleta” fuera sujeto y “tío” fuera objeto. La asociación semántica A -sujeto tiene relación con B -objeto, es correcta cuando A =tío, B =bicicleta, pero no lo es en caso contrario. Como podemos ver, A y B están vinculados mediante una relación R , que tiene sentido si la tripleta (A,R,B) existe, pero (B,R,A) no existe.

El problema que nos concierne en esta tesis es estudiar la forma en que podemos modelar relaciones entre palabras, y más generalmente, relaciones entre entidades de texto, lo que algoritmos como Word2Vec no pueden hacer de manera directa.

De esta forma, tenemos tres tipos de relación entre palabras que a continuación definiremos.

Al observar una relación R para dos pares de palabras (por ejemplo, $R = \text{amod}$ de la Tabla 1.3) se define un conjunto E (por ejemplo $\{\text{cuchillo}, \text{cortar}, \text{afilado}\}$ de la Tabla 1.3) de entidades que están involucradas con la relación R . Usando esta notación, definimos tres tipos de relaciones:

- R es una relación de tipo asimétrica si para todo par de palabra w_1, w_2 en E si se cumple que w_1 está relacionado con w_2 mediante R , entonces w_2 no se debe relacionar con w_1 a través de R (en esta dirección). Formalmente, $\forall x, y \in A : xRy$, entonces yRx no existe.
- R es una relación de tipo simétrica si para todo par de palabra w_1, w_2 en E si se cumple que w_1 está relacionado con w_2 mediante R , entonces w_2 también está relacionado con w_1 a través de R (en esta dirección). Formalmente, $\forall x, y \in A : xRy \rightarrow yRx$.
- R es una relación de tipo antisimétrica si para todo par de palabra w_1, w_2 en E si se cumple que w_1 está relacionado con w_2 mediante R y w_2 también está relacionado con w_1 a través de R , en esta dirección, entonces w_1 es igual a w_2 . Formalmente, $\forall x, y \in A : xRy \wedge yRx \rightarrow x = y$.

Como se puede observar en la Tabla 1.2 el modelo Word2Vec solo es capaz de modelar relaciones de tipo simétricas. Sin embargo, en el lenguaje natural las palabras no guardan relaciones de tipo simétricas entre sí. Retomando el ejemplo (cuchillo, amod, afilado) de la Tabla 2, si la relación amod fuese de tipo simétrica entonces afilado se relacionaría con cuchillo a través de amod (en esta dirección) lo cual es falso porque “cuchillo” no es adjetivo y “afilado” no es sustantivo. De esta forma podemos concluir que la relación amod es de tipo asimétrica.

A este estudio de relaciones entre palabras se le conoce como Modelado de relaciones, el cual se describe a continuación.

1.2.1. Modelado de relaciones

Dado un conjunto de tripletas, por ejemplo Tabla 1.3, definido como Ω , se construyen los siguientes conjuntos.

- El conjunto de entidades, definido por $E = \{e_1, e_2, \dots, e_n\}$. Es decir, los sujetos y objetos únicos en Ω . Por ejemplo, de la Tabla 1.3 $E = \{\text{cuchillo}, \text{cortar}, \text{afilado}\}$.
- El conjunto de relaciones, definido por $R = \{r_1, r_2, \dots, r_m\}$. Es decir, las relaciones únicas existentes en Ω . Por ejemplo, de la Tabla 1.3 $R = \{\text{amod}, \text{obj}\}$.

A partir de esto, se construye un tensor de tercer orden X que modela m relaciones entre n entidades, tal que $m = |R|$ y $n = |E|$, de la siguiente forma.

Para cada relación $r \in R$, se contruye una matriz X_r de tamaño $n \times n$, donde las filas y columnas corresponden a la estidades en E , tal que cada entrada (i,j) de la matriz X_r es un valor del conjunto $\{-1, 1\}$. Es decir,

$$X_r = \begin{cases} 1 & \text{si } (e_i, r, e_j) \in \Omega \\ -1 & \text{en otro caso} \end{cases},$$

donde e_i y e_j son entidades sujeto y objeto, respectivamente, existentes en E .

Por ejemplo, la matriz X_{amod} modela la relación “amod” entre las entidades $E = \{cuchillo, cortar, afilado\}$ extraídas de la Tabla 1.3.

$$X_{amod} = \begin{bmatrix} -1 & -1 & 1 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Ante este estudio de Modelado de relaciones han surgido modelos capaces de crear representaciones incrustadas de palabras usando el modelado de relaciones en su forma tensorial.

A continuación, se mencionan 3 modelos.

1.3. Estado del arte

El primer modelo [7] a describir, está basado en el marco de la Memoria Distribuida [8](MD por sus siglas). El marco de la MD representa la distribución de la información en un corpus como un tensor de tercer orden, es decir, un objeto ternario que toma la forma de tripletas (sujeto, relación, objeto) a la que se le asigna un peso w que codifica el número de coocurrencias entre palabras. En la Tabla 1.4 se pueden observar unos ejemplos. Por ejemplo, el tensor del MD contiene la tripleta $(knife, nmod - 1, sharp)$ lo que indica que la palabra “sharp” ocurre en el corpus como adjetivo de “knife” con un peso de 156.63. Éste peso nos indica que “sharp” es el adjetivo que más coocurre con el sustantivo “knife”.

Subject	Relation	Object	w
knife	nmod	sharp	156.63
knife	is	sharp	80.07
knife	nmod	blunt	69.66

Tabla 1.4: Ejemplos de tripletas MD para el sustantivo knife

El enfoque propuesto es construir incrustaciones de palabras de la siguiente forma. Sean N el vocabulario de sustantivos y A en vocabulario de adjetivos extraídos del tensor MD, con $n_A = |A|$. La j -ésima palabra en el vocabulario de sustantivos se representada por un vector con entradas reales y de dimensión n_A , construido de la siguiente forma.

$$e_j = \begin{bmatrix} adj_{1,j} \\ adj_{2,j} \\ \vdots \\ adj_{n_A,j} \end{bmatrix}$$

Donde $adj_{i,j}$ es calculado a partir del peso w asociado a la tripleta $(j, nmod, i)$, que corresponde al sustantivo $j \in E$ y al adjetivo $i \in A$ del tensor MD.

Por ejemplo, “knife” es la entrada 10343 del vocabulario de sustantivos y una visualización parcial de su incrustación se muestra a continuación, donde al lado derecho se muestran algunos adjetivos relacionados con la palabra “knife” [7].

$$e_{10343} = \begin{bmatrix} 0 \\ 4.329 \\ 16.368 \\ \vdots \\ 73.297 \\ \vdots \\ 327.629 \\ \vdots \\ 0 \end{bmatrix} \begin{array}{l} \text{Aboriginal-j} \\ \text{Afghan-j} \\ \text{African-j} \\ \vdots \\ \text{long-j} \\ \vdots \\ \text{sharp-j} \\ \vdots \\ \text{zealous-j} \end{array}$$

De aquí se puede observar que “sharp” es el adjetivo que tiene mayor influencia en la representación vectorial del sustantivo “knife”. Esto porque la entrada correspondiente a “sharp” en el vector e_{10343} es el máximo..

La construcción de las incrustaciones de palabras es un procedimiento que se realiza con el tensor MD, y se ejecuta en tres pasos:

1. Extracción de una lista de sustantivos.
2. Extracción de una lista de adjetivos.
3. Construcción de la matriz de incrustación.

Este modelo basado en MD goza de las propiedades de las incrustaciones de palabras aprendidas, y al mismo tiempo su proceso de construcción es comprensible, a diferencia de modelos que utilizan Redes Neuronales Artificiales. Sin embargo, solo modela relaciones entre palabras de tipo $nmod$ u otro, pero solo una relación a la vez.

El segundo modelo por presentar es el modelo RESCAL [9] el cual es un modelo de datos multirrelacionales. El enfoque propuesto es a través de una factorización tensorial de rango r que tiene en cuenta la estructura de los datos relacionales.

Formalmente, dado un conjunto de tripletas Ω con su respectivo conjunto de entidades \mathcal{E} y relaciones \mathcal{R} . Se construye el tensor X que modela m relaciones entre n entidades existentes en Ω , con $m = |\mathcal{R}|$ y $n = |\mathcal{E}|$. Cada matriz X_k del tensor X se factoriza como:

$$X_k = AR_kA^T, \text{ para } k = 1, \dots, m.$$

Donde A es una matriz de tamaño $n \times r$ que contiene las representaciones de los componentes latentes de las entidades en \mathcal{E} . La matriz R_k es una matriz asimétrica que modela las interacciones de los componentes latentes en la k -ésima relación en \mathcal{R} .

Las matrices A y R_k se calculan resolviendo el problema de minimización con regularización

$$\min_{A, R_k} f(A, R_k) + g(A, R_k) \quad (1.1)$$

donde

$$f(A, R_k) = \frac{1}{2} \sum_k (\|X_k - AR_kA^T\|^2) \quad (1.2)$$

y el término de regularización g es

$$g(A, R_k) = \frac{1}{2} \lambda (\|A\|^2 + \sum_k \|R_k\|^2). \quad (1.3)$$

Este modelo a diferencia del anterior no limita a un vocabulario, tampoco se limita a un solo tipo de relación. Sin embargo, por el teorema espectral para matrices ortogonalmente diagonalizables, sabemos que cada matriz X_k es simétrica, lo cual implica que este modelo sólo modela relaciones simétricas. Dado que el interés del presente proyecto es modelar relaciones asimétricas o antisimétricas, el modelo RESCAL no es adecuado.

El tercer modelo, el modelo de estudio del presente proyecto, es el modelo ComplEx [10] que a continuación se describe.

Dado un conjunto de tripletas Ω con su conjunto de entidades E y conjunto de relaciones R . Se construye el tensor X que modela m relaciones entre n entidades existentes en Ω . En este contexto, el objetivo del modelo ComplEx es determinar si una tripleta $(s, r, o) \in \Omega$ existe. Es decir, predecir la entrada $X_{s,r,o}$ correspondiente a la tripleta (s, r, o) a través de una función de puntuación definida como ϕ .

ComplEx propone una factorización tensorial de componentes latentes de un tensor, utilizando la diagonalización unitaria para la factorización. La cual se describe con mayor detalle en la Sección 2.1.

El modelo ComplEx fue propuesto para crear representaciones incrustadas de las entidades en E y las relaciones en R , creando representaciones en el dominio complejo. Al utilizar la factorización en

el dominio complejo, el modelo ComplEx es capaz de factorizar matrices que modelan relaciones de tipo asimétricas (por ejemplo, relaciones gramaticales entre palabras) o incluso antisimétricas. A diferencia del modelo RESCAL el cual solo factoriza matrices que modelan relaciones simétricas. Debido a que las relaciones simétricas y antisimétricas son el interés del presente proyecto, se eligió ComplEx para el estudio de las relaciones discursivas.

En este contexto, ahora se describe la hipótesis que se plantea en esta tesis.

1.4. Pregunta de investigación

¿Puede el modelo ComplEx modelar relaciones lingüísticas de un grado de abstracción mayor que el derivado de coocurrencias léxicas, como relaciones léxico-gramaticales y semánticas?

1.5. Hipótesis

Si podemos encontrar una forma de representar relaciones discursivas, entonces el modelo ComplEx nos permitirá analizar geoméricamente estas relaciones.

1.6. Objetivos

A continuación se describen los objetivos que se pretenden lograr en el presente proyecto.

1.6.1. Objetivo general

- Estudiar las representaciones generadas por ComplEx para relaciones discursivas.

1.6.2. Objetivos específicos

- Modelación de rasgos lingüísticos para relaciones discursivas.
- Implementación del modelo ComplEx.
- Analizar las representaciones construidas por el modelo.

2.1. Modelo de estudio: ComplEx

Una matriz real cuadrada X se factoriza como $EW E^T$, esta factorización es usada en matrices simétricas reales, las cuales modelan relaciones simétricas. Ya que el interés de este proyecto es modelar relaciones de tipo asimétricas y antisimétricas, la factorización de matrices reales simétricas no es útil en este caso. Por lo cual se necesita una factorización de matrices cuadradas reales utilizando la diagonalización unitaria, una generalización de la diagonalización ortogonal para matrices complejas. Esta generalización permite una factorización de matrices cuadradas reales arbitrarias, es decir, matrices simétricas, asimétricas y antisimétricas.

Una matriz $X \in \mathbb{C}^{n \times n}$ es unitariamente diagonalizable si y solo si puede ser escrita como

$$X = EWE^*,$$

donde $E, W \in \mathbb{C}^{n \times n}$ tal que W es diagonal y E es una matriz unitaria. Una matriz es unitaria si cumple que la matriz por su traspuesta es igual a la matriz identidad.

El Teorema Espectral para matrices normales dice que una matriz X cuadrada compleja es unitariamente diagonalizable si y solo si X es normal. Por lo cual todas las matrices reales son normales, más aún el conjunto de todas las matrices normales incluye todas las matrices antisimétricas reales y muchas otras matrices que son útiles para representar diferentes tipos de relaciones, por ejemplo relaciones asimétricas.

Se generaliza, mostrando que para cualquier $X \in \mathbb{R}^{n \times n}$ existe una diagonalización unitaria en el dominio complejo cuya parte real es igual a X .

$$X = \text{Re}(EWE^*)$$

En otras palabras, la diagonalización unitaria se proyecta sobre el subespacio real.

El modelo ComplEx [10] hace uso de esta factorización, por lo que es capaz de modelar relaciones simétricas, antisimétricas o asimétricas. El enfoque de este modelo es el siguiente.

Dado un conjunto de tripletas Ω con su respectivo conjunto de entidades \mathcal{E} y relaciones \mathcal{R} . Se construye el tensor X que modela m relaciones entre n entidades existentes en Ω , con $m = |\mathcal{R}|$ y $n = |\mathcal{E}|$.

Cada matriz X_r del tensor X , se factoriza como

$$X_r = \text{Re}(EW_r E^*), \text{ donde } r = 1, \dots, m, \text{ tal que}$$

- La matriz $E \in \mathbb{C}^{n \times k}$ contiene las representaciones incrustadas de las entidades. Es decir, la i -ésima fila de E corresponde a la i -ésima entidad en \mathcal{E} .

- Cada matriz $W_r \in \mathbb{C}^{k \times k}$ representa la r -ésima relación en \mathcal{R} .

De esta forma, la función de puntuación ϕ para la tripleta $(s, r, o) \in \Omega$ queda de a siguiente forma.

$$\phi(s, r, o) = \text{Re}(\langle w_r, e_s, e_o \rangle), \text{ donde} \quad (2.1)$$

- Los vectores $e_s, e_o \in \mathbb{C}^k$ son las representaciones incrustadas de las entidades s y o , respectivamente. También, son filas de la matriz E .
- El vector $w_r = \text{diag}(W_r) \in \mathbb{C}^{n \times k}$ es la representación incrustada de la r -ésima relación en \mathcal{R} .
- El valor k es un hiperparámetro de rango, es la dimensión de incrustación para las entidades y relaciones.

Los parámetros a aprender, denotado por Θ , en el modelo ComplEx son

$$\Theta = \{e_i, w_r \in \mathbb{C}^k | i \in \mathcal{E} \text{ y } r \in \mathcal{R}\}.$$

Desarrollando la ecuación 2.1 se tiene

$$\begin{aligned} \phi(s, r, o) = & \langle \text{Re}(w_r), \text{Re}(e_s), \text{Re}(e_o) \rangle \\ & + \langle \text{Re}(w_r), \text{Im}(e_s), \text{Im}(e_o) \rangle \\ & + \langle \text{Im}(w_r), \text{Re}(e_s), \text{Im}(e_o) \rangle \\ & - \langle \text{Im}(w_r), \text{Im}(e_s), \text{Re}(e_o) \rangle. \end{aligned}$$

De esta forma la función de puntuación ϕ solo involucra vectores reales correspondientes a las partes reales e imaginarias de las incrustaciones de entidades y relaciones. El conjunto de parámetros Θ queda de la siguiente forma:

$$\Theta = \{e'_i, e''_i, w'_r, w''_r \in \mathbb{R}^k | i \in \mathcal{E} \text{ y } r \in \mathcal{R}\},$$

donde $e'_i = \text{Re}(e_i)$, $e''_i = \text{Im}(e_i)$, $w'_r = \text{Re}(w_r)$ y $w''_r = \text{Im}(w_r)$.

Los parámetros Θ se calculan resolviendo un problema de optimización, utilizando el descenso de gradiente estocástico. Las actualizaciones de la tasa de aprendizaje se realizan con el algoritmo AdaGrad [11]. Optimizando la probabilidad logarítmica negativa de la función ϕ con una regularización L2 en los parámetros. La optimización es aplicada a cada tripleta $(r, s, o) \in \Omega$ y su etiqueta y , que corresponde a la entrada $X_{r,s,o}$ en el tensor que modela las relaciones. La función a optimizar es

$$\gamma(r, s, o; \Theta) = \sum_{((r;s;o);y)} \log(1 + e^{-y\phi(r,s,o;\Theta)}) + \lambda \|\Theta\|, \quad (2.2)$$

donde $\lambda \in \mathbb{R}^+$ es el parámetro de regularización.

2.2. Red semántica WordNet

WordNet [12] codifica relaciones semánticas entre palabras mediante un grafo. Usa las categorías sintácticas sustantivo, verbo, adjetivo y adverbio. WordNet incluye las siguientes relaciones semánticas:

- Sinonimia: Relación básica en WordNet, utiliza conjuntos de sinónimos (synsets) para representar los sentidos de las palabras.
- Antonimia: Relación entre formas de palabras, es útil para organizar los significados de adjetivos y adverbios.
- Hiponimia e hiperonimia: Incluye relaciones transitivas entre synsets, es útil para organizar los significados de los sustantivos en una estructura jerárquica.
- Meronimia y holonimia: Relaciones semánticas complejas. WordNet distingue partes componentes, partes sustantivos y partes de miembros.

WordNet también es conocido como un grafo de conocimiento. Este grafo de conocimiento no está completo, ya que existen muchas relaciones entre palabras faltantes. Con la finalidad de completar este tipo de grafos surge el Aprendizaje Relacional Estadístico, cuyo objetivo es predecir las relaciones faltantes con el aprendizaje de las relaciones existentes. El modelo RESCAL y el modelo ComplEx han sido entrenados con la red WordNet para realizar puntos de referencia.

2.3. Técnica de reducción de dimensionalidad: t-SNE

El algoritmo t-distributed Stochastic Neighbor [13] (t-SNE) es un método de reducción de dimensionalidad, utilizado principalmente para la visualización de datos.

Sea X el conjunto de puntos, tal que $X_i \in \mathbb{R}^n$, donde $i = 1, \dots, m$. El algoritmo se divide en 3 etapas, que son las siguientes:

1. La primera etapa del algoritmo es calcular las distancias euclidianas entre cada par de puntos (X_i, X_j) para $i = 1, \dots, m$ y $j = i+1, \dots, m$. Tomando estas distancias y transformándolas en probabilidades condicionales que representan la similitud entre cada punto. Esto para evaluar qué tan similares son cada dos puntos en los datos, en otras palabras, qué tan probable es que sean vecinos.

La probabilidad condicional de que el punto X_j este cercano al punto X_i está representada por un gaussiano centrado en X_i . Formalmente, la probabilidad $p_{j|i}$ se define como:

$$p_{j|i} = \frac{e^{-\|X_i - X_j\|^2 / 2\sigma_i^2}}{\sum_{k \neq i} e^{-\|X_i - X_k\|^2 / 2\sigma_i^2}}$$

A partir de las probabilidades condicionales creadas, calculamos la Distribución de Probabilidad Conjunta P , de la siguiente forma:

$$P_{j|i} = \frac{p_{j|i} + p_{i|j}}{2m}$$

2. En la segunda etapa se crean los datos en una dimensión reducida.

Construimos el conjunto Y de puntos aleatorios con la misma cantidad de puntos que en existen el conjunto original, donde cada punto esta en \mathbb{R}^k , donde k es la nueva dimensión y calculamos también una Distribución de Probabilidad Conjunta para estos nuevos puntos definida a partir de las probabilidades condicionales $q_{j|i}$.

La probabilidad condicional $q_{j|i}$ que se define para estos nuevos datos es mediante la distribución t (de Student), es de la siguiente forma:

$$q_{ij} = \frac{e^{-\|Y_i - Y_j\|^2}}{\sum_{k \neq i} e^{-\|Y_i - Y_k\|^2}}$$

3. En la tercera etapa se corrigen los punto en Y de modo que su Distribución de Probabilidad Conjunta sea lo más similar posible a la de los datos originales. Esto se realiza utilizando el Descenso de Gradiente Estocástico. La función de error que el Descenso de Gradiente intenta minimizar es la divergencia Kullback-Leiber (divergencia KL) de la distribución de probabilidad conjunta P del espacio original y Q del espacio de baja dimensión. La función de costo se define de la siguiente forma:

$$C = \sum_i KL(P_i|Q_i) = \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

A partir de esta optimización, obtenemos el conjunto Y , donde $Y_i \in \mathbb{R}^k$.

3.1. Corpus de Relaciones discursivas

El corpus utilizado es el Penn Discourse Treebank-3 [14] el cual es un corpus a gran escala anotado con información relacionada con la estructura y semántica del discurso. El PDTB-3 se enfoca en codificar las relaciones del discurso, la metodología de anotación utilizada sigue un enfoque basado en el léxico con el fin de permitir que el corpus se utilice dentro de diferentes marcos teóricos. Cada relación discursiva en el corpus se anota independiente de otras relaciones, es decir, la dependencia entre relaciones no está marcada.

Se considera que en una relación discursiva existen dos segmentos de discurso (párrafos, oraciones o frases subordinadas) que están relacionados por un conectivo explícito (ver ejemplo 1) ó implícito (ver ejemplo 2).

En general, cada documento en el corpus consta al menos de los siguientes tres elementos: El primer argumento de la relación (*Arg1*) escrito en cursiva; el conectivo o marcador en la relación discursiva, que está subrayado y viene acompañado de información entre paréntesis sobre el tipo (explícito o implícito) y sentido de la relación; el segundo argumento (*Arg2*) escrito en negrita.

Ejemplo (1) *The Manhattan U.S. attorney's office stressed criminal cases from 1980 to 1987, averaging 43 for every 10,000 adults.*

(Explicit, Contrast)

But **the New Jersey U.S. attorney average 16.**

Ejemplo (2) *So far, the mega-issues are a hit*

(Implicit, Instantiation, For example)

Earlier this year, Tata Iron & Steel Co.'s offer of \$ 355 million of convertible debentures was over subscribed.

En el caso de que el marcador discursivo sea de tipo implícito en una relación, los anotadores agregaron en paréntesis el marcador discursivo más adecuado para la relación discursiva, como en el ejemplo 2 el marcador es *For example*. El PDTB-3 posee 29818 relaciones discursivas; combinando relaciones discursivas con marcadores explícitos e implícitos. Las relaciones discursivas están divididas por categorías. En la Figura 3.1 se puede observar la jerarquía de la categoría Comparación. Las relaciones discursivas utilizadas en el presente proyecto son de categoría Comparación, específicamente las de tipo contraste.

A continuación, vamos a describir el proceso que seguimos para alcanzar nuestros objetivos. Lo primero es representar cada documento de manera adecuada, para poder extraer información sobre las relaciones discursivas de interés. Para ello, por cada documento se extraen los tres elementos antes mencionados y se forma una tripleta (*Arg1*, *conectivo*, *Arg2*). La Tabla 3.1 muestra ejemplos

Sense
Comparison.Concession+SpeechAct.Arg2-as-denier+SpeechAct
Comparison.Concession.Arg1-as-denier
Comparison.Concession.Arg2-as-denier
Comparison.Contrast
Comparison.Similarity Expansion.Manner.Arg2-as-manner
Comparison.Similarity
Contingency.Cause+Belief.Reason+Belief
Contingency.Cause+Belief.Result+Belief
Contingency.Cause.Reason Expansion.Instantiation.Arg2-as-instance
Contingency.Cause.Reason Expansion.Level-of-detail.Arg2-as-detail
Contingency.Cause.Reason
Contingency.Cause.Result Expansion.Manner.Arg1-as-manner
Contingency.Cause.Result
Contingency.Condition.Arg1-as-cond
Contingency.Condition.Arg2-as-cond
Contingency.Negative-condition.Arg2-as-negCond
Contingency.Purpose.Arg1-as-goal
Contingency.Purpose.Arg2-as-goal Expansion.Manner.Arg1-as-manner
Contingency.Purpose.Arg2-as-goal

Figura 3.1: Ejemplo de los niveles (separados por un punto) para la categoría Comparación de tripletas.

<i>Arg1</i>	<i>conectivo</i>	<i>Arg2</i>
The Manhattan U.S. ... adults	but	the New Jersey ... 16
...
So far, the ... hit	For example	Earlier this year, ... subdescribed

Tabla 3.1: Conjunto de relaciones discursivas

Una vez realizado el proceso anterior, se observaron las siguientes características en el conjunto de tripletas:

- Se obtuvieron 1702 tripletas.
- Los marcadores discursivos encontrados son: *however, although, but, nevertheless, while, whereas, by contrast, in comparison, by comparison, in contrast, in fact, as, on the contrary, yet, if, and, on the other hand y still*.
- Para una tripleta (*Arg1, conectivo, Arg2*), el *Arg1* nunca se repite en otra tripleta y el *Arg2* produce el mismo fenómeno.
- Para una tripleta (*Arg1, conectivo, Arg2*), el *Arg1* nunca aparece en otra tripleta como argumento 2, y viceversa.

Unos de los objetivos de este proyecto es observar si existen patrones que nos permitan saber distinguir un marcado discursivo de otro. Por ejemplo, diferenciar patrones entre relaciones discursivas con marcador “but” y “on the other hand”. Si trabajamos con las tripletas tal y como

aparecen en la Tabla 3.1, difícilmente podremos extraer información útil, ya que como se puede apreciar, en el lenguaje natural hay muy pocos patrones que podríamos identificar a nivel léxico. Debido a esto, se propuso un método de representación para argumentos el cual se describe en la siguiente sección.

3.2. Señales léxicas

Se realizó la creación de una serie de conjuntos, elegidos a partir de 4 rasgos lingüísticos (señales léxicas) específicamente el tipo de palabra o la categoría gramatical de la palabra, de la siguiente forma.

Definimos a V como el vocabulario de los argumentos Arg1 y Arg2 en todas las tripletas de la Tabla 3.1. Para cada palabra $w \in V$ tal que w es un adjetivo, sustantivo, verbo o adverbio creamos el conjunto

$$C_i = \left\{ \begin{array}{c|c} \textit{sinónimos} & \textit{antónimos} \\ \hline \textit{de} & \textit{de} \\ \textit{w} & \textit{w} \end{array} \right\}$$

Para obtener los sinónimos de w se utilizó la relación de sinonimia que WordNet posee. Para obtener los antónimos se utilizó la relación de antonimia en WordNet. Ejemplos de estos conjuntos se pueden observar en la Figura 3.2.

$$\left\{ \begin{array}{c|c} \textit{larg} & \textit{small} \\ \textit{big} & \textit{little} \\ \textit{heavy} & \dots \\ \dots & \dots \end{array} \right\} \quad \left\{ \begin{array}{c|c} \textit{profit} & \\ \textit{benefit} & \textit{loss} \\ \textit{hit} & \\ \dots & \dots \end{array} \right\} \quad \left\{ \begin{array}{c|c} \textit{complete} & \\ \textit{end} & \textit{begin} \\ \textit{stop} & \\ \dots & \dots \end{array} \right\} \quad \left\{ \begin{array}{c|c} \textit{always} & \\ \textit{ever} & \textit{never} \\ \textit{forever} & \end{array} \right\}$$

(a) Conjunto de ad- (b) Conjunto de (c) Conjunto de ver- (d) Conjunto de ad-
jetivos sustantivos bos verbios

Figura 3.2: Ejemplos de conjuntos por señales léxicas.

Los conjuntos tienen la siguiente característica: si w es un adjetivo entonces la parte izquierda y la parte derecha del conjunto C_i solo contiene adjetivos, si w fue un sustantivo entonces el conjunto C_i contiene palabras que son sustantivos; la misma característica se refleja para el caso de las señales verbo y adverbio. Después de la creación de los conjuntos se obtuvo lo siguiente:

- Se obtuvieron 1413 conjuntos.
- Cada conjunto contiene al menos un elemento en su lado izquierdo y en su lado derecho.
- El conteo de conjuntos por señales léxicas es el siguiente: son 248 conjuntos de adjetivos, 531 conjuntos de sustantivos, 588 conjuntos de verbos y 46 conjuntos de adverbios.

Lo siguiente que se realizó, fue un ordenamiento de los conjuntos de acuerdo a su POS.

- Para cada tripleta $(R_{Arg1}, conn, R_{Arg2})$ en Ω , R_{Arg1} se repite en otra tripleta como R_{Arg2} y viceversa.

Como se mencionó en la sección 3.1, el objetivo es observar patrones que nos permitan distinguir marcadores discursivos de otros. Por lo cual se propuso este método de representación a través de las señales léxicas adjetivo, sustantivo, verbo y adverbio en el texto. Esto con el fin de comprobar si estas señales son adecuadas para realizar distinción entre relaciones discursivas con diferentes marcadores discursivos.

3.4. Relación semántica: sinonimia-antonimia

El segundo rasgo que se estudió en las relaciones discursivas, es la relación semántica de sinonimia-antonimia entre argumentos de una tripleta en el conjunto Ω (definido en la sección anterior). La relación semántica se estudio de la siguiente forma.

Para cada tripleta $(Arg1, conn, Arg2)$ tenemos una representación rala de los argumentos $Arg1$ y $Arg2$ la cual puede ser vista como un vector de 1413 entradas, por ejemplo

$$\begin{aligned} Arg1 &= (-1, 0, 1, 0, \dots, 0, -1, 1, 0) \\ Arg2 &= (1, -1, 1, 0, \dots, 0, 0, -1, 0). \end{aligned}$$

Para detectar la sinonimia y antonimia en la tripleta realizamos el producto elemento a elemento, por ejemplo

$$Arg1 * Arg2 = (-1, 0, 1, 0, \dots, 0, 0, -1, 0).$$

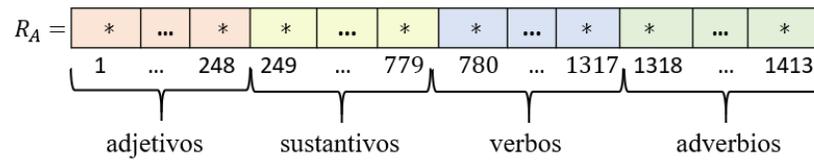
Del cual obtenemos lo siguiente:

- Contamos la cantidad de “1” en el resultado del producto, y lo definimos como n_{sin} . Este número nos indicará si existe sinonimia entre $Arg1$, $Arg2$ y el grado en que existe. En el ejemplo anterior, $n_{sin} = 1$ el cual indica que sí existe sinonimia entre los argumentos y el grado es 1.
- Contamos la cantidad de “-1” en el resultado del producto, y lo definimos como n_{ant} . Este número nos indicará si existe antonimia entre $Arg1$, $Arg2$ y el grado. En el ejemplo, $n_{sin} = 2$ el cual indica que sí existe antonimia y el grado es 2.

De esta forma, a la tripleta $(Arg1, conn, Arg2)$ le asignamos una representación bidimensional (n_{sin}, n_{ant}) en términos de sinonimia y antonimia. Se propuso este segundo rasgo para ver si es posible observar diferentes patrones en las relaciones discursivas por marcador discursivo.

3.5. Conjunto de entrenamiento para cada señal léxica

Las señales léxicas (tipo o categoría gramatical de una palabra) utilizadas son: adjetivo, sustantivo, verbo y adverbio. Cada argumento A tiene una representación de longitud 1413, esta representación se puede dividir en 4 partes debido al ordenamiento realizado en los conjuntos (definidos en la sección 3.2) de acuerdo con su POS; esta división se observa de la siguiente forma.



Con esta división realizada, se crearon 4 modelos entrenados independientemente, mediante 4 conjuntos de entrenamiento, usando las representaciones de cada señal léxica, ver Figura 3.3.

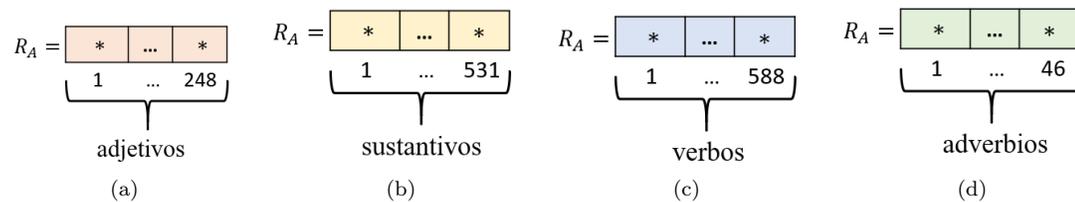


Figura 3.3: Representaciones por señal.

3.6. Propiedades geométricas de las incrustaciones

Dada la tripleta (s, r, o) queremos saber si se satisface que

$$vec(s) + vec(r) \approx vec(o). \tag{3.1}$$

Esta operación implica que, si tenemos el vector sujeto, el vector relación y el vector objeto (ejemplo en la Figura 3.4(a) para el caso de dos dimensiones). Entonces la proyección del vector sujeto sobre el vector relación y el vector objeto sobre el vector relación aparecen en cierto orden con respecto al vector relación, como se observa con el punto naranja y azul de la Figura 3.4(a). Por lo tanto, para saber si se satisface la ecuación 3.1 se necesita observar el orden en que están las proyecciones; si los puntos naranjas aparecen después de los puntos azules, entonces la ecuación 3.1 se satisface. Por ejemplo, las nubes de puntos de las proyecciones podrían tener el orden mostrado en la imagen 3.4(b), en este caso la ecuación 3.1 se satisface. Se realizó un conjunto de proyecciones para todas las tripletas, separando datos por marcadores discursivos y aplicado a cada modelo de las señales léxicas.

Se utilizaron las siguientes medidas de calidad, para observar que tan bien estas proyecciones producen el orden deseado.

- **Proyección positiva:** como el punto naranja (proyección del objeto sobre la relación) debe quedar después del punto azul (proyección del sujeto sobre la relación). Para comprobar si el orden de las proyecciones es el correcto hacemos lo siguiente. Por cada tripleta, obtenemos la proyección naranja y azul. Restamos el valor de la proyección naranja al valor de la proyección azul. Si la resta es negativa, entonces el orden de la proyección es el correcto. Hacemos esto con todas las tripletas y calculamos la proporción de las restas que dan resultado negativo. Definimos esta proporción como *Proy. Pos.*

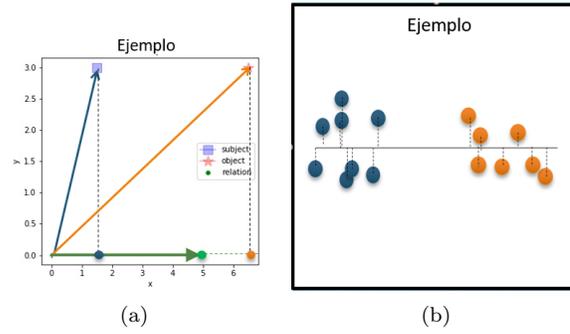


Figura 3.4: Ejemplos de las propiedades geométricas.

- Error promedio: Se define como el promedio de las distancias entre $vec(s) + vec(r)$ y $vec(o)$, calculadas por cada tripleta utilizando la distancia angular. Esta medida nos indica cual es el error en la ecuación 3.1. Definimos este promedio de distancias como Err. Prom.

Resultados

4.1. Relaciones discursivas

Se realizó un estudio para determinar qué tipo de relación (simétrica, asimétrica, antisimétrica) es cada marcador discursivo para cada representación realizada con las señales léxicas. Esto con el fin de determinar si estas representaciones de argumentos son útiles para utilizar el modelo ComplEx. Recordemos que este modelo es útil para relaciones que son de tipo simétricas, asimétricas y antisimétricas.

El resultado se colocó en la figura 4.1. Podemos observar que la mayoría de los marcadores dis-

Relación	Rep. Adjetivos			Rep. Sustantivos			Rep. Verbos			Rep. Adverbios		
	Simetría	Asimetría	Antisimetría	Simetría	Asimetría	Antisimetría	Simetría	Asimetría	Antisimetría	Simetría	Asimetría	Antisimetría
by contrast	0	95	5	0	88	12	0	96	4	0	90	10
but	2	94	4	0	95	5	0	90	10	20	73	7
whereas	0	96	4	0	85	15	0	88	12	0	100	0
though	0	100	0	---	---	---	---	---	---	---	---	---
with	0	100	0	---	---	---	---	---	---	---	---	---
however	0	97	3	---	---	---	---	---	---	0	85	15
in contrast	0	100	0	0	92	8	0	88	12	0	75	25
nevertheless	0	100	0	0	88	12	0	100	0	0	100	0
by comparison	0	96	4	0	100	0	0	80	20	0	91	9
still	0	100	0	0	91	9	0	100	0	0	100	0
although	0	87	13	0	100	0	0	100	0	0	100	0
on the contrary	0	100	0	0	100	0	0	50	50	---	---	---
while	0	94	6	0	75	25	0	75	25	19	66	15
on the other hand	0	94	6	0	86	14	0	90	10	0	100	0
in fact	0	100	0	0	90	10	---	---	---	0	100	0
and	0	100	0	0	100	0	0	100	0	0	100	0
conversely	0	100	0	0	50	50	0	100	0	---	---	---
if	0	0	100	---	---	---	---	---	---	---	---	---
regardless	0	100	0	0	100	0	---	---	---	---	---	---
yet	0	100	0	---	---	---	0	100	0	---	---	---
in comparison	0	100	0	0	100	0	0	100	0	0	100	0
as	0	0	100	0	100	0	---	---	---	---	---	---
meanwhile	---	---	---	0	100	0	---	---	---	---	---	---

Figura 4.1: Resumen de tipo de relación por marcador discursivo y señal léxica

cursivos son de tipo asimétricos, esto para cada representación realizada con las señales. Con este resultado se comprobó que el método de representación de argumentos propuesto en el presente proyecto es útil para utilizar el modelo ComplEx ya que fue propuesto para aprovechar este tipo de relaciones, a diferencia del modelo RESCAL y el modelo Word2Vec. En el caso de las representaciones realizadas por las señales sustantivo y adjetivo se observó que algunas tripletas con diferentes marcadores discursivos cumplen con las características de tipo antisimétrica.

4.2. Visualización de las propiedades geométricas

Los conjuntos de proyecciones, uno por cada modelo de señal, se graficaron en una gráfica de tipo violintplot. Se observaron tres fenómenos principales, los cuales son los siguientes.

- El primer fenómeno observado es el que corresponde con la relación deseable, donde las proyecciones están en el orden correcto. Por ejemplo, el primer violín de la Figura 4.2(a) para la relación “although” en el modelo de señal verbo muestra un ejemplo de este primer fenómeno. En este ejemplo, la medida Proy. Pos. nos indica que el 100 % de las proyecciones quedan en el orden correcto.
- El segundo fenómeno, el cual predomina más, es en el que las proyecciones quedan muy juntas o incluso pueden estar encimados puntos azules y naranjas. Por ejemplo, el segundo violín de la Figura 4.2(a) para la relación “while” en el modelo de señal verbo. En este ejemplo, la medida Proy. Pos. indica que el 66 % de las proyecciones están en el orden correcto.
- El tercer fenómeno, el fenómeno no deseado, es el fenómeno inverso del primero. Por ejemplo, el tercer violín de la Figura 4.2(a) para la relación “still” en el modelo de señal verbo. En este ejemplo, la medida Proy. Pos. indica que el 40 % de las proyecciones están en el orden correcto.

La medida Err. Prom, en los tres fenómenos nos va indicando que: para el primer fenómeno el error es pequeño; para el segundo fenómeno el error es significativamente mayor al primer fenómeno; para el tercero y último fenómeno, en algunos casos el error se observa marcadamente menor al del primer fenómeno, ya que éste es inverso de aquél.

Para los modelos restantes de las señales adverbio, sustantivo y adjetivo se muestran estos tres fenómenos en la Figura 4.2(b), 4.2(c) y 4.2(d) respectivamente.

4.3. Señales léxicas

En cada modelo entrenado, uno por cada señal, se observaron diferentes comportamientos en el resultado de las incrustaciones. Por ejemplo, el comportamiento de las entidades involucradas con el marcador discursivo “on the other hand” para el modelo de la señal verbo se ve en la Figura 4.3(a). El comportamiento de las entidades con respecto al mismo marcador discursivo para el modelo de la señal adverbio se ve en la Figura 4.3(b). En este caso, para las señales verbo y adverbio, se notó un comportamiento similar. Esto debido que en ambos conjuntos de entrenamiento el conjunto de entidades; es decir, los argumentos 1’s y los argumentos 2’s únicos, contienen muy pocos. En ambos casos, los sujetos ó argumentos 1’s y los objetos ó argumentos 2’s se repiten con mucha frecuencia cumpliendo el mismo rol.

El comportamiento de las entidades involucradas con el marcador discursivo “on the other hand” para el modelo de la señal sustantivo se ve en la Figura 4.4(a). El comportamiento de las entidades con respecto al mismo marcador discursivo para el modelo de la señal adjetivo se ve en la Figura 4.4(b).

Finalmente, aquí se observó que las incrustaciones de entidades para las señales sustantivo y adjetivo producen un comportamiento similar. En este caso, en ambos conjuntos de entrenamiento, las entidades no se repiten con frecuencia. También, en comparación a las señales anteriores,

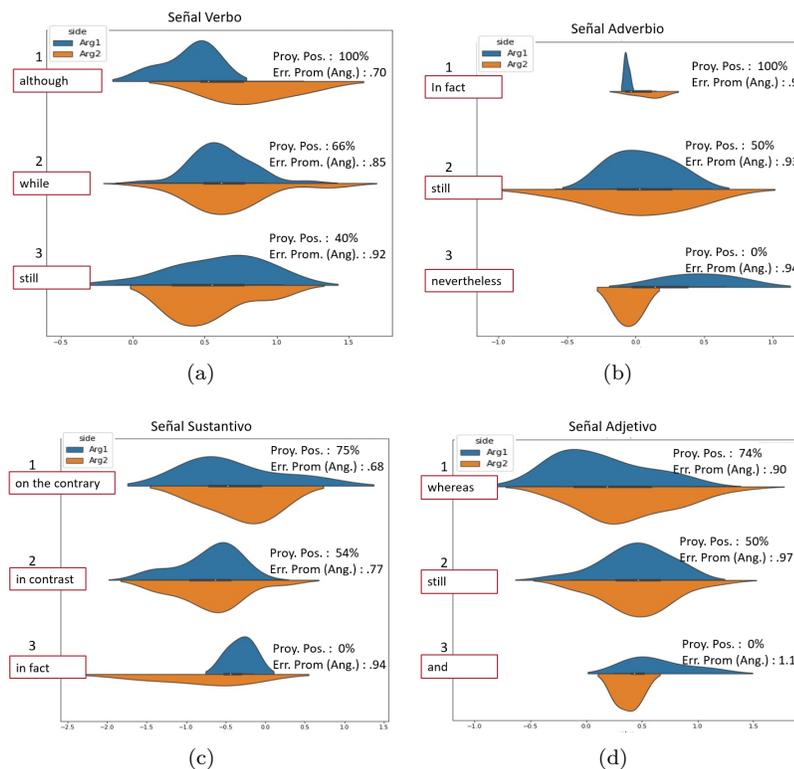


Figura 4.2: Gráficas de violín utilizando proyecciones, descritas en la Sección 3.6.

los sujetos aparecen rara vez como objetos y los objetos como sujetos para este marcador discursivo.

Debido a que se observó un comportamiento diferente en los cuatro modelos, uno por cada señal léxica, pero a la vez se observó que las entidades para los modelos con señales verbo y adverbio producen un comportamiento similar. También, en los modelos de señales sustantivo y adjetivo producen un comportamiento similar. Se realizó una comparativa de sus vectores relación, es decir, se graficaron las incrustaciones de los marcadores discursivos para modelos que corresponden a la señal verbo y adverbio en el mismo espacio, esto con la aplicación del algoritmo de reducción de dimensionalidad t-SNE. El resultado se observa en la Figura 4.5(a), en el cual se produjo una separación de marcadores discursivos por tipo de señal, de modo que fue posible entrenar un modelo de clasificación no-lineal para visualizar la frontera de separación.

Se realizó la misma comparativa para los modelos de las señales sustantivo y adjetivo. El resultado se puede observar en la Figura 4.5(b). Nuevamente, los marcadores se separaron por tipo de señal, en este caso se entrenó un modelo de clasificación lineal, en la gráfica se observa la frontera resultante que separa los marcadores discursivos por tipo de señal.

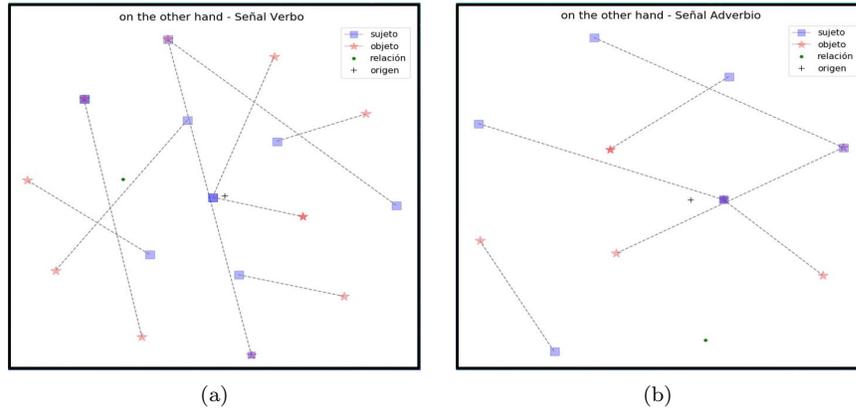


Figura 4.3: Incrustaciones de entidades para el marcador “on the other hand” de las señales verbo y adverbio.

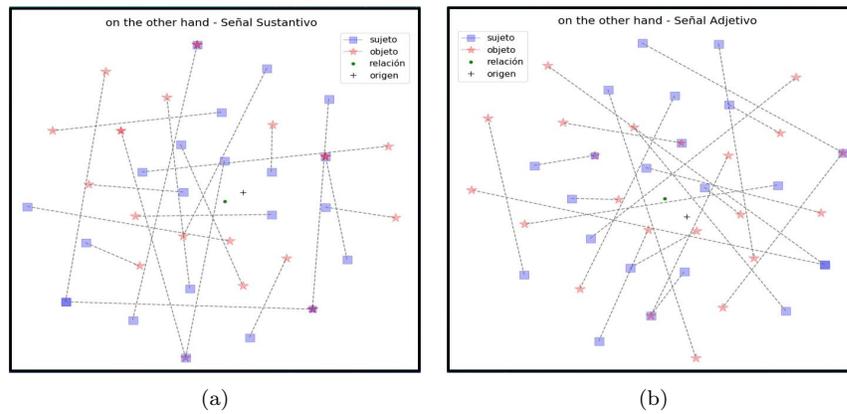


Figura 4.4: Incrustaciones de entidades para el marcador “on the other hand” (señales sustantivo y adjetivo).

4.4. Señales semánticas: sinonimia-antonimia

Los resultados obtenidos con respecto a las señales semánticas sinonimia y antonimia, descrita en la sección 3.4, se describen a continuación. Se encontraron 4 principales fenómenos que predominan en las relaciones discursivas de categoría Comparación en el nivel de contraste, por marcador discursivo.

- Se encontraron relaciones discursivas con cierto tipo de marcador discursivo en las cuales solo se expresa la señal sinonimia. Por ejemplo, para relaciones discursivas con marcador “although”, ver Figura 4.6(a).

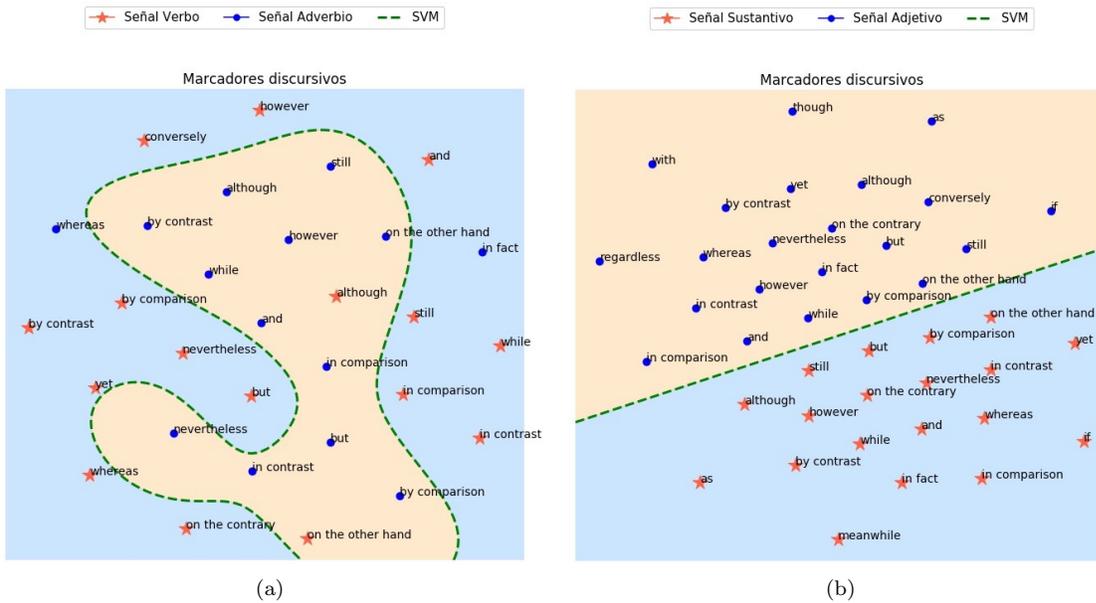


Figura 4.5: Incrustaciones de marcadores discursivos.

- Se encontraron relaciones discursivas en las cuales solo se expresa la señal antonimia. Por ejemplo, para relaciones discursivas con marcador “meanwhile”, ver Figura 4.6(b).
- Relaciones discursivas en las cuales se expresa una combinación de sinonimia y antonimia. Por ejemplo, para relaciones discursivas con marcador “however”, ver Figura 4.6(c).
- Relaciones discursivas en las cuales no se expresaron las señales semánticas. Por ejemplo, para relaciones discursivas con marcador “like”, ver Figura 4.6(d).

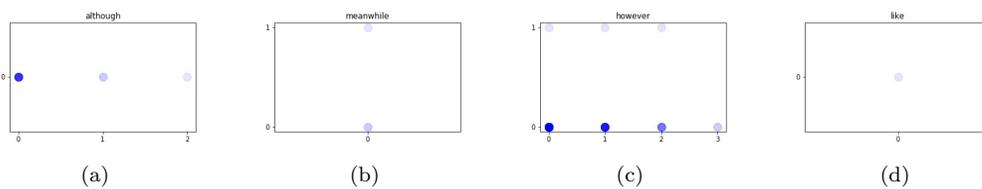


Figura 4.6: Gráfica de los puntos (n_{ant}, n_{sin}) por marcador discursivo. Mientras más fuerte sea el color más puntos caen en la misma posición.

Como se mencionó en la Sección 3.1, los marcadores discursivos pueden aparecer de forma explícita o implícita en las relaciones discursivas. Se realizó un estudio de estas señales semánticas

separando relaciones discursivas por marcadores explícitos o implícitos el cual se condensa en el mapa de calor de la Figura 4.7(a). En esta figura se observa que en la mayoría de las relaciones

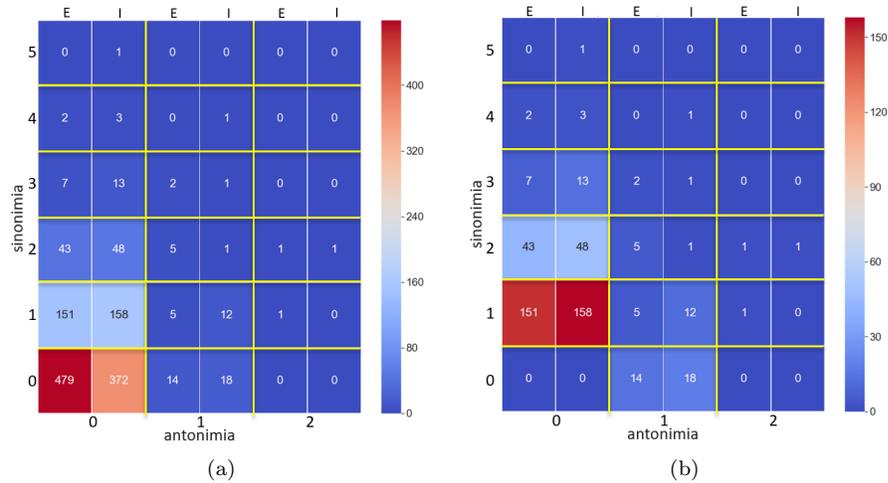


Figura 4.7: Mapa de calor del conteo de puntos (n_{ant} , n_{sin}) separando marcadores explícitos e implícitos. Cada columna está marcada por tipo explícito (E) o implícito (I).

discursivas, para marcadores explícitos e implícitos, no se expresa la sinonimia y la antonimia. Para el caso de relaciones discursivas con marcadores explícitos existen 479 en las que no se expresan las señales y para marcadores implícitos existen 372.

Sin embargo, existe una proporción considerable de relaciones discursivas en las cuales sí se expresan las señales semánticas para marcadores explícitos e implícitos. Existen 151 relaciones discursivas con marcadores explícitos en las que se expresa la sinonimia con grado 1 y 43 con grado 2. Existen 158 relaciones discursivas con marcadores implícitos en las que se expresa la sinonimia con grado 1 y 48 con grado 2.

Con respecto a la señal antonimia, existen 14 relaciones discursivas con marcadores explícitos en las que se expresa la antonimia con grado 1 y 18 relaciones discursivas con marcadores implícitos. La cantidad de relaciones discursivas en las que se expresan las señales semánticas se puede observar con mayor claridad en la Figura 4.7(b), en la cual se colapsaron a 0 todos los conteos de las relaciones discursivas en las que no se expresa la sinonimia y antonimia para marcadores explícitos e implícitos.

CAPÍTULO 5

Conclusiones

El método de representación de argumentos propuesto en el presente proyecto, además de crear una representación vectorial de textos también codifica dentro información de las señales semánticas sinonimia-antonimia, esto por la naturaleza del proceso.

Como se mencionó, en la construcción de las representaciones de argumentos para cada relación discursiva, estas pueden ser divididas en las cuatro señales léxicas. Lo cual nos proporciona cuatro representaciones de un argumentos, uno por cada señal léxica, esto nos permitió analizar las relaciones discursivas separándolas por señal para ver que señal aporta más a la representación creada y así de ser posible, formular una representación final quitando o agregar señales léxicas.

Se automatizó la representación de estos rasgos lingüísticos en cada una de las tripletas. La hipótesis es que si estas señales semánticas son importantes entonces la representación obtenida tendrá un comportamiento distinto para cada marcador discursivo.

Se observó que nuestras representaciones sí permiten distinguir un conector de otro, así como establecer una diferencia entre marcadores explícitos e implícitos.

Sin embargo, por una parte, parece que no son señales abundantes en relaciones con este sentido de contraste, y por otra parte, parece que las representaciones obtenidas no se comportan como se esperaba. Esto último podría deberse a que la forma en como construimos nuestras entidades no es la adecuada, o quizá, a que el algoritmo que elegimos no sea el más adecuado para estudiar estas señales lingüísticas.

El análisis usando diferentes rasgos lingüísticos, permite observar cuáles de estos contribuyen en mayor medida a la “señalización” de las relaciones de contraste mediante señales semánticas de sinonimia-antonimia.

Finalmente, se observó que los sustantivos y adjetivos son los que contribuyen en mayor medida a la representación de relaciones de contraste, mientras que verbos y adverbios, a pesar de ser “palabras de contenido” o palabras conceptuales parecen tener una menor influencia.

Como trabajo futuro se pretende explorar nuevas técnicas de representaciones de texto para ser aplicadas a los argumentos de las relaciones discursivas. Otra posible ruta a seguir, es realizar una adaptación del modelo ComplEx matizar la escala o tipo de relaciones existentes, y así tener más de dos clases en cada modelado matricial.

Bibliografía

- [1] William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175. Citeseer, 1994.
- [2] Christopher Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.
- [3] Zhang Yun-tao, Gong Ling, and Wang Yong-cheng. An improved tf-idf approach for text classification. *Journal of Zhejiang University-Science A*, 6(1):49–55, 2005.
- [4] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.
- [5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [6] Owen Rambow, K Vijay-Shanker, and David Weir. D-tree grammars. *arXiv preprint cmp-lg/9505028*, 1995.
- [7] Lauro Snidaro, Giovanni Ferrin, and Gian Luca Foresti. Distributional memory explainable word embeddings in continuous space. In *2019 22th International Conference on Information Fusion (FUSION)*, pages 1–7. IEEE, 2019.
- [8] Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.
- [9] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Icml*, volume 11, pages 809–816, 2011.
- [10] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. *International Conference on Machine Learning (ICML)*, 2016.
- [11] Agnes Lydia and Sagayaraj Francis. Adagrad—an optimizer for stochastic gradient descent. *Int. J. Inf. Comput. Sci.*, 6(5), 2019.
- [12] Christiane Fellbaum. Wordnet. *The encyclopedia of applied linguistics*, 2012.
- [13] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [14] Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. The penn discourse treebank 3.0 annotation manual, 2019.