**Bloque II: Homogeneización y refinamiento**

## **Este bloque aborda las limitaciones de OpenIE relacionadas con la delimitación incorrecta en las extracciones y el fenómeno de la extracción poco informativa, en el que las tuplas generadas omiten información clave. El objetivo de este conjunto de estrategias es refinar los predicados mal extraídos y unificar las entidades que hacen referencia a un mismo concepto.**

## **Estrategia 3:** Corrección de verbos auxiliares y modales

La hipótesis inicial planteaba que, al incorporar los modales (“can”, “may”, “should”, “will”, etc.) en el propio predicado y limpiar al mismo tiempo los objetos, se obtendrían relaciones con predicados más informativos y objetos más “puros”, lo que debería aumentar la precisión del grafo.

Al aplicar esta transformación a los textos de **Utilitarianism** y **Ellipse**, los únicos en los que se habían detectado duplicaciones de verbos modales, se observa que predicados genéricos como “can” pasan a formar parte de su complemento. Por ejemplo, “can describe by” sustituye al predicado genérico “can”, y el objeto “can describe by set of point in coordinate plane” se reduce a “set of point in coordinate plane”. De igual modo, “can discern throughout history of ethical theory” se separa en el predicado “can discern throughout” y el objeto “history of ethical theory”. En los textos **Homeostasis** e **Inflation** no hay tripletas con modales, por lo que sus métricas quedarán idénticas a las de la Transformación 1.

**Análisis por texto**

[Poner imágenes B3U, B3E]

* **Utilitarianism**: En el caso de **Utilitarianism**, como muestra la figura X, ambas precisiones experimentan una ligera mejora. La cobertura semántica permanece constante, mientras que la cobertura léxica destaca por aumentar un 9,3 %. Este incremento probablemente se deba al proceso de limpieza de predicados y objetos. En conjunto, el aumento en estas métricas refleja una mejora en la calidad del grafo.
* **Ellipse**: En este caso las mejoras son aún más destacables. Las precisiones semántica y léxica mejoran un **x %** y un **x %** respectivamente. Además, la limpieza de verbos y objetos parece haber mejorado también las coberturas en este caso. El aumento de las métricas de Jaccard, refleja también una mejora clara en la estructura del grafo.

En conjunto, la corrección de los verbos modales ha limpiado de manera efectiva las tripletas afectadas en los textos de Utilitarianism y Ellipse, lo que se traduce en un aumento notable tanto de la precisión como de la cobertura del grafo.

**Evolución media de métricas**

[Poner imagen L3]  
En la media de las métricas de Utilitarianism y Ellipse (figura X) se aprecia con claridad que la Transformación 3 mejora todas las métricas. Esto confirma la hipótesis de que, al depurar los verbos modales y limpiar los objetos, se obtienen predicados más completos y objetos más precisos, lo que incrementa la precisión del grafo y su calidad general.

## **Estrategia 4:** Mapeo de sujetos y objetos a chunks de CoreNLP

La hipótesis de la Estrategia 4 plantea que, al reemplazar en el grafo generado por OpenIE las entidades originales por sintagmas nominales más limpios y gramaticalmente correctos, se obtendría un grafo más limpio y más preciso.

Al aplicar el mapeo descrito en la metodología se obtienen ejemplos exitosos como los siguientes:

* (position, articulate until, century) → (utilitarian positions, articulate until, the 19th century)
* (increase, be in, demand pull inflation) → (an increase, be in, demand - pull inflation)

Sin embargo, para comprobar si esta estrategia realmente mejora los resultados es imprescindible evaluar su impacto a partir de las métricas correspondientes.

**Análisis por texto**

[Poner imágenes B4U, B4E, B4H, B4I]

**Utilitarianism:** En el caso de Utilitarismo, se puede observar que la aplicación de la Estrategia 4 sí mejora los resultados de forma general. Las métricas semánticas experimentan un aumento más notable, mientras que las métricas léxicas se mantienen prácticamente constantes. Lo más destacado es el incremento en la cobertura semántica, posiblemente debido a que el mapeo de chunks ha permitido que aparezcan en el grafo sintagmas más completos que antes no se capturaban.

**Ellipse:** En el caso de **Ellipse**, los resultados empeoran. Todas las precisiones y coberturas caen de forma notable respecto al grafo con la Transformación 1, especialmente las métricas léxicas. Esto indica que los chunks extraídos por CoreNLP y utilizados para el mapeo son menos precisos que las entidades identificadas originalmente por OpenIE.

**Homeostasis:** Para el texto de la Homeostasis se aprecia una clara diferencia entre las métricas semánticas y las léxicas. Por ejemplo, mientras que la precisión semántica mejora en un x%, la léxica empeora un x%. Esto indica que, aunque las entidades generadas ahora están más alejadas en superficie del grafo gold, en realidad se acercan más a nivel de significado. Dicho de otro modo, los chunks extraídos por CoreNLP, a pesar de a escribirse de forma distinta, comparten un sentido más parecido con los nodos del grafo gold.

**Inflation:** En el caso de Inflación se aprecian mejoras, aunque bastante leves. La única métrica que disminuye es la cobertura léxica, lo que indica que los nuevos nodos están algo más alejados en su forma escrita del grafo gold. Sin embargo, en términos generales, la Estrategia 4 ha aportado una ligera mejora global para el texto sobre la inflación.

En conjunto, aunque la Estrategia 4 muestra potencial, en algunos casos como el de elipse, los resultados empeoran. Esto sugiere que sería necesario mejorar la calidad de los chunks utilizados en el mapeo, para que representen de forma más fiel las entidades de cada frase y permitan mapear sujetos y objetos con mayor precisión.

**Evolución media de métricas**

[Poner imagen L4]

En la figura X se aprecia que únicamente las métricas semánticas mejoran al aplicar la Estrategia 4. La precisión semántica experimenta un incremento del X%, mientras que la precisión léxica empeora un X%. De igual modo, la cobertura semántica sube un X%, pero la cobertura léxica desciende un X%. En cuanto a la similitud estructural, el coeficiente de Jaccard semántico aumenta un X%, mientras que el Jaccard léxico disminuye un X%.

Los resultados indican que, aunque la estrategia depura el grafo desde el punto de vista semántico, no mejora su calidad léxica. Por tanto, convendría optimizar tanto la extracción de chunks con CoreNLP como el propio proceso de mapeo, de modo que los sujetos y objetos resultantes reflejen con mayor fidelidad las entidades originales del texto.

## Estrategia 5: Mapeo de sujetos y objetos con chunks limpiados y enriquecidos (CoreNLP + SpaCy)

La hipótesis señalaba que, al mejorar y enriquecer la calidad de los chunks y al flexibilizar el proceso de mapeo, se aumentaría el número de asignaciones correctas de sujetos y objetos, traduciéndose en un grafo más limpio y preciso.

La estrategia 5 es similar a la 4, pero trtando de mejorar los chunks usados en la estrategia anterior y flexibilizando el mapeo para mejorar los resultados. Omo se explica en la metodología, se tratrá da superar las carencias de la genereacion de chunks de la estrategia anterior mediante la limpieza de stop words iniciales de los chunks, la eliminación de chunks que contengan verbos y añadir los sintagmas nominales extraídos por SpaCy.

La Estrategia 5 retoma la Estrategia 4 pero incorpora mejoras en la calidad de los chunks y flexibiliza el mapeo para optimizar los resultados. En primer lugar, se eliminan las stop words iniciales de cada chunk. Después, se descartan aquellos chunks que contengan verbos para preservar únicamente sintagmas nominales. Finalmente, se enriquecen los candidatos añadiendo los sintagmas nominales extraídos por SpaCy.

**Análisis por texto**

[Poner imágenes B5U, B5E, B5H, B5I]

**Utilitarianism:** Se aprecia una mejora en todas las métricas. Al compararlo con los avances que ofrecía la estrategia anterior en este texto del dataset de referencia, se observa que ahora la mejora es más notable. Especialmente en el ámbito léxico, donde con la estrategia previa las métricas se mantenían relativamente constantes frente al grafo de la Transformación 1.

**Ellipse:** Mientras que con el mapeo de la estrategia anterior casi todas las métricas empeoraban, ahora con los nuevos chunks y el mapeo flexible se puede observar una gran mejora. La cobertura léxica es la única métrica que no mejora respecto al grafo obtenido en el bloque anterior. Aun así, su descenso es menor que el experimentado con la Estrategia 4. Esto sugiere que los nuevos chunks limpios y enriquecidos son más cercanos a los del grafo gold, y que además están llevando a más entidades a ser mapeadas a chunks precisos y coherentes.

**Homeostasis:** En este caso se aprecia de nuevo una mejora general, sobre todo a nivel semántico, donde los avances son más pronunciados. Sin embargo, a diferencia de la Estrategia 4 en la que todas las métricas léxicas empeoraban de forma notable, con la nueva estrategia solo disminuye la cobertura léxica, y lo hace en una magnitud mucho menor (con la Estrategia 4 caía un X %, mientras que ahora solo lo hace un Y %).

**Inflation:** Nuevamente, mientras con la Estrategia 4 las métricas se mantenían casi constantes, ahora se aprecia una mejora destacada, con un incremento superior al X % en las precisiones. Las coberturas apenas varían, lo que sugiere que la extracción de chunks previa ya capturaba adecuadamente las entidades del texto. La similitud estructural con el grafo gold también mejora: el coeficiente de Jaccard semántico sube un X % y el léxico aumenta un Y %.

En este análisis por texto se evidencia de nuevo que mapear entidades a chunks no solo mejora la precisión del grafo, sino que también amplía su cobertura. Esto puede deberse a que muchas de las entidades extraídas por OpenIE son menos precisas que emplear directamente los sintagmas nominales detectados en la frase.

También se observa que, aunque esta estrategia puede disminuir ligeramente la cobertura sobretodo léxica en algunos casos, la mejora sustancial en el resto de las métricas compensa con creces.

Esta pequeña caída en la cobertura puede explicarse porque ahora existen menos formas de referirse a un mismo concepto: dos expresiones que antes se representaban con ligeras variaciones léxicas se unifican en un solo mapeo. Es posible que, al hacerlo, la entidad aparezca con una forma menos parecida en superficie al nodo correspondiente del grafo gold, pero al mismo tiempo se evita tener dos nodos que representen la misma entidad, lo cual mejora la consistencia y claridad global del grafo.

**Evolución media de métricas**

[Poner imagen L5]

En la figura X se aprecia que la Estrategia 5 mejora de manera global la calidad del grafo, superando con creces los resultados obtenidos con la Estrategia 4.

Todas las métricas experimentan mejoras de entre un 10 % y un 35 %, salvo la cobertura léxica, que se reduce en menos de un 1 %. Como se explicó anteriormente, esta ligera caída puede deberse a la desaparición de algunas variantes léxicas de las entidades a causa del mapeo. En conjunto, estos resultados confirman que, al mejorar y enriquecer la calidad de los chunks y al flexibilizar el proceso de mapeo, se obtiene un grafo más limpio y preciso, sobre todo a nivel semántico.

Evaluación del bloque

El objetivo de este bloque era unificar entidades equivalentes y refinar predicados mal extraídos, garantizando una representación más coherente y clara de las relaciones.

En primer lugar, los resultados de la Estrategia 3, diseñada para corregir los verbos modales que provocan extracciones erróneas de tripletas por parte de OpenIE, confirman que esta intervención incrementa la precisión y la calidad general del grafo.

Esto indica que la Estrategia 3 será beneficiosa en los casos en que los verbos modales dificulten la extracción de tripletas coherentes, cumpliendo así parte del objetivo de este segundo bloque de estrategias. Por tanto, se decide implementarla de aquí en adelante.

En segundo lugar, tras aplicar la Estrategia 3, disponemos de dos enfoques para el mapeo a chunks: un método más estricto que utiliza directamente los chunks de CoreNLP (Estrategia 4) o un enfoque más flexible que emplea los chunks previamente limpiados y enriquecidos con los sintagmas nominales extraídos por spaCy.

En la figura X puede verse que la Estrategia 5 ofrece la mayor mejora global de resultados, a cambio de una ligera reducción de la cobertura léxica respecto al grafo obtenido tras aplicar las mejoras del Bloque 1.

[Poner imagen evB2B.png]

Por tanto, en este bloque se optará por las Estrategias 3 y 5, ya que son las que mejor cumplen el objetivo de unificar entidades equivalentes y depurar predicados mal extraídos, logrando así una mejora significativa de la calidad general del grafo.