UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MIARFID Official Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

Master Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital
Universidad Politécnica de Valencia

Resúmenes Automáticos Extractivos y Abstractivos

TRABAJO TECNOLOGÍAS DEL LENGUAJE HUMANO

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

André Pachedo,
Miquel Gómez

CAPÍTULO 1
Análisis de los Resultados

En esta memoria se responde a las preguntas planteadas en la Práctica de Tecnologías del Lenguaje Humano (TLH) sobre los resultados obtenidos en el notebook sobre *automatic summarization*.

Se han utilizado LLMs para la interpretación inicial de los resultados con el objetivo de tener una primera idea. Luego, se han analizado de forma detallada y se han llegado a conclusiones propias. La redacción se ha hecho a mano, usando estos modelos como apoyo en la corrección del texto y el formato.

1.1 Análisis de Rendimiento

Primero que todo, basándonos en las tablas de resultados, podemos decir que los oráculos extractivos, marcan el máximo valor para las distintas métricas que los métodos estudiados pueden alcanzar para cada dataset. A partir de estos valores, podremos saber cuanto margen de mejora tiene cada método.

Por otra parte, el método *naive* LEAD-3 nos permite caracterizar los datasets, indicando cuánta información relevante se encuentra en las tres primeras frases de los documentos.

Cabe destacar que, por definición, ningún método extractivo puede superar las métricas de los oráculos, ya que estos seleccionan la mejor combinación de frases basándose directamente en la referencia.

1.1.1. Dataset CNN/DailyMail

Tabla 1.1: Resultados de métricas automáticas para modelos extractivos y abstractivos (Tabla 1)

Method	R-1	R-2	R-3	R-4	BS-r	BS-f	METEOR	CHRF
extractive-oracle-num-sents	**0.4617**	**0.2840**	**0.2155**	**0.1711**	**0.8792**	**0.8847**	**0.4165**	**47.7765**
extractive-oracle-ratio	**0.4646**	**0.2829**	**0.2126**	**0.1666**	**0.8834**	**0.8860**	**0.4371**	**50.7997**
lead-3	0.3788	0.1629	0.0931	0.0633	0.8695	0.8666	0.3683	42.8866
tex-trank	0.3459	0.1526	0.0905	0.0637	0.8726	0.8625	0.3561	41.9320
lex-trank	0.3473	0.1386	0.0764	0.0514	0.8672	0.8635	0.3301	39.4642
lsa	0.3123	0.1196	0.0690	0.0470	0.8641	0.8540	0.3274	39.1290
bert-extractive	0.3321	0.1212	0.0649	0.0440	0.8637	0.8596	0.3184	38.7921
bart	**0.4298**	0.2031	0.1231	0.0848	0.8778	**0.8804**	**0.3880**	**44.1239**
pegasus	0.4209	**0.2115**	**0.1297**	**0.0896**	**0.8827**	0.8780	0.3580	41.6167
flan-t5	0.3567	0.1481	0.0796	0.0490	0.8580	0.8689	0.2811	34.4151
gpt2	0.1970	0.0322	0.0059	0.0017	0.8106	0.8149	0.1759	26.7276
llama3	0.2287	0.0475	0.0135	0.0046	0.8384	0.8413	0.2150	30.0470

Los métodos extractivos (TextRank, LexRank, LSA, BERT) muestran un comportamiento estable. Obtienen resultados competentes en todas las métricas, aunque no destacan en ninguno y tienen aún margen de mejora. Vemos que donde más fallan es en la coherencia, la consistencia tampoco es muy alta, exceptuando a TextRank, cuyo valor en BertScore-R es competitivo. Luego, para la fluidez, en comparación con los modelos abstractivos, los extractivos quedan claramente por detrás, mientras que en Relevancia tienen valores decentes.

En cuanto a los modelos *Sequence2Sequence* abstractivos, observamos una clara distinción. BART y Pegasus ofrecen los mejores resultados en todas las métricas, distanciándose de los métodos extractivos en varias de ellas. Por ejemplo, hay una mejora notable en la coherencia, destacando Bart en la métrica de CHRF, obteniendo resultados cercanos al del oráculo. La consistencia también la mejoran considerablemente, llegando a mejorar Pegasus en BertScore-R al primer oráculo basado en el número de oraciones. Lo mismo ocurre con la fluidez, donde Bart consigue acercarse mucho al oráculo en la métrica de METEOR. Sin embargo, donde más sobresalen estos dos métodos es en la Relevancia, donde todas las métricas relacionadas con este aspecto (BertScore-F, METEOR y CHRF), reciben una mejora sustancial.

Por el contrario, Flan-T5 presenta un rendimiento inferior a estos dos, más cercano a los modelos extractivos. La razón de esto radica en que BART y Pegasus han sido finetuneados en el dataset CNN/DailyMail, mientras que para Flan-T5 se desconoce si es así. En todas las métricas se queda cerca de los extractivos pero sin llegar a superarlos en ninguna, estando de forma consistente por debajo de ellos.

Finalmente, los modelos causales, GPT-2 y Llama-3, exhiben el rendimiento más bajo en las métricas automáticas. Como se profundizará en la sección de impacto de la abstractividad, esto no implica que sus resúmenes sean de mala calidad. La causa principal es su naturaleza generativa, ya que estos modelos tienden a parafrasear y utilizar 'sus propias palabras' mucho más que los modelos *Sequence2Sequence*. Dado que la mayoría de métricas están basadas en comparaciones de n-gramas con la referencia, estos modelos son altamente penalizados al introducir nuevas formas de expresar la misma idea. Por todo esto, en los cuatro aspectos de Coherencia, Consistencia, Fluidez y Relevancia, los modelos causales quedan muy por debajo de los demás métodos.

Viendo los resultados globales, podemos decir que el método extractivo más balanceado es TextRank, mientras que BART destaca como el modelo abstractivo más equilibrado, empatando con Pegasus en varias métricas y manteniendo una alta consistencia.

1.1.2. Dataset CLASum

Tabla 1.2: Resultados de métricas automáticas para modelos extractivos y abstractivos (Tabla 2)

Method	R-1	R-2	R-3	R-4	BS-r	BS-f	METEOR	CHRF
extractive-oracle-num-sents	**0.4254**	**0.2731**	**0.2184**	**0.1840**	**0.8815**	**0.8870**	**0.3908**	**48.1411**
extractive-oracle-ratio	**0.4303**	**0.2779**	**0.2229**	**0.1872**	**0.8853**	**0.8881**	**0.4109**	**51.4000**
lead-3	0.3576	0.1615	0.1097	0.0871	0.8772	0.8711	0.3419	**42.4826**
tex-trank	0.3187	0.1333	0.0861	0.0661	0.8754	0.8640	0.3398	39.5359
lex-trank	0.3521	0.1530	0.0975	0.0734	0.8782	0.8710	0.3384	41.1056
lsa	0.3102	0.1256	0.0806	0.0621	0.8733	0.8608	0.3325	39.2861
bert-extractive	0.3218	0.1228	0.0746	0.0549	0.8711	0.8631	0.3154	37.9083
bart	0.3783	0.1818	0.1244	0.0984	0.8753	0.8770	0.3218	41.0940
pegasus	0.3798	0.1904	0.1305	**0.1012**	0.8788	0.8712	0.3232	40.5537
flan-t5	**0.4187**	**0.2074**	**0.1341**	0.0992	**0.8793**	**0.8884**	**0.3348**	40.2826
gpt2	0.1870	0.0314	0.0063	0.0009	0.7380	0.7424	0.1466	22.9283
llama3	0.2325	0.0500	0.0135	0.0042	0.7876	0.7927	0.1959	27.8445

En este dataset, el panorama cambia significativamente, ya que ningún modelo ha sido entrenado específicamente en CLASum, por lo que los resultados reflejan mejor la capacidad de generalización de cada método.

Respecto a los métodos extractivos, aunque el rendimiento general empeora ligeramente, se mantiene la tendencia observada anteriormente, pero con LexRank destacando por encima del resto.

El cambio más relevante ocurre en los modelos *Sequence2Sequence*. Flan-T5 obtiene los mejores resultados en casi todas las métricas, superando a BART y Pegasus. Ahora, viendo los valores, las diferencias no son tan pronunciadas como en el dataset anterior, por lo que no se puede menospreciar la capacidad de generalización de BART y Pegasus. Todos obtienen resultados competitivos, pero Flan-T5 consigue los mejores valores en Relevancia (BertScore-F, METEOR) y en Consistencia (R-3, BertScore-R), mientras que Pegasus gana en Coherencia y Fluidez (R-4 y CHRF).

Finalmente, los modelos causales (GPT2 y Llama3) tienen resultados similares a los del dataset anterior. Siguen teniendo un rendimiento inferior al de los demás métodos, aunque en este caso, Llama-3 obtiene mejores resultados que GPT2 en todas las métricas. Por lo mencionado anteriormente, siguen en desventaja debido a su naturaleza generativa.

1.1.3. Comentario sobre los oráculos

En ambos datasets, los métodos que no son muy abstractivos (extractivos y *Sequence2Sequence*) se acercan bastante a los oráculos, especialmente en las métricas de Relevancia y Consistencia. También, si nos fijamos en el método naive LEAD-3, vemos que obtiene resultados decentes e incluso mejores que los demás métodos.

Con todo esto, llegamos a la conclusión de que ambos datasets tienen un nivel de abstractividad bajo, y que los resúmenes son en gran parte extractivos. Los oráculos marcan un límite alto, y el hecho de que los métodos extractivos y *Sequence2Sequence* (que no son muy abstractivos) se acerquen tanto a ellos confirma esta teoría. Los resultados del LEAD-3 también apoyan esta idea.

Por lo tanto, los dos métodos más abstractivos (modelos causales) están en clara desventaja, por todo lo comentado anteriormente.

1.2 Análisis de Comparativo

Comparando el comportamiento de los modelos entre los distintos datasets, observamos patrones muy diferenciados. Los métodos extractivos demuestran una gran estabilidad en todos los aspectos, manteniendo métricas consistentes tanto en CNN/DailyMail como en CLASum. Por el contrario, los modelos abstractivos Sequence2Sequence, BART y Pegasus, empeoran bastante del primer corpus al segundo, su rendimiento en la relevancia y coherencia disminuye notablemente. Ahora, el modelo Flan-T5, que no había sido finetuneado en ninguno de estos datasets, mejora en CLASum. Por último, los modelos causales GPT-2 y Llama-3 mantienen un rendimiento bajo en ambos datasets, con ligeras variaciones pero sin cambios significativos.

En este punto, si tuviéramos que elegir un método para trabajar en ambos datasets, elegiríamos un método abstractivo basado en modelos Sequence2Sequence. Entre ellos, es complicado decidir cuál de los tres, sin embargo, nos quedaríamos con Pegasus.

En el primer conjunto de datos, Pegasus y BART tenían un rendimiento muy bueno por su ventaja de haber sido entrenados específicamente para este dataset. Luego, en este segundo dataset, Flan-T5 ha tenido un rendimiento muy bueno, pero Pegasus no se ha quedado muy atrás. Este modelo ha obtenido resultados muy buenos en ambos datasets. A pesar de que no ha sido el mejor en ninguno, ha demostrado ser un modelo robusto y consistente. Parece ser el más fácil de adaptar a distintos tipos de documentos y tareas en términos generales.

Además, hay que tener en cuenta los resúmenes que generan los modelos causales, que con las métricas usadas, que se apoyan en referencias extractivas, no se les está haciendo justicia. Por lo que habría que revisar las producciones de estos modelos mediante el juicio humano u otra métrica para valorar su calidad real.

1.3 Impacto de la Abstractividad

Tabla 1.3: Correlaciones de Kendall Tau entre el rendimiento global y el nivel de abstractividad

Method	Type	Aspect	Abs-Level Correlation
extractive-oracle-num-sents	extractive	performance	-0.625105
extractive-oracle-ratio	extractive	performance	-0.632635
lead-3	extractive	performance	-0.614855
tex-trank	extractive	performance	-0.528799
lex-trank	extractive	performance	-0.503332
lsa	extractive	performance	-0.476109
bert-extractive	extractive	performance	-0.514631
bart	abstractive	performance	-0.543865
pegasus	abstractive	performance	-0.581181
flan-t5	abstractive	performance	-0.450436
gpt2	abstractive	performance	-0.064108
llama3	abstractive	performance	-0.135576

Viendo la tabla con las correlaciones de "Kendall tau", se aprecia que existe una correlación negativa evidente entre el rendimiento de los métodos y el nivel de abstractividad. Como se ha apuntado durante el análisis, esto es debido al diseño de las métricas automáticas, que se basan en referencias concretas, y que en este trabajo parecen ser extractivas. Si los resúmenes de referencia se vuelven muy abstractos, se alejan del texto fuente, lo que penaliza a cualquier modelo que se ciña al texto original.

Todos los métodos extractivos y los *Sequence2Sequence*, presentan una correlación negativa con valor absoluto notable, mientras que los causales tienen una correlación cercana a cero.

Tal como se adelantó en el análisis de rendimiento, los modelos no causales, dan resumenes bastante ligados al texto original y son poco, sino nada, capaces de abstraer. Esto les deja con una relación clara entre el nivel de abstractividad del resumen y su rendimiento, ya que si los resúmenes de referencia son muy abstractivos, estos modelos no podrán acercarse a ellos.

Por otro lado, gpt2 y llama3, al ser modelos generativos, tienen la capacidad de abstraer y parafrasear, por lo que no se ven afectados por el nivel de abstractividad de los resúmenes de referencia. Ahora, no hay que olvidar su rendimiento global es bajo, de por si no conseguen grandes puntuaciones en las métricas.