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

Los métodos generativos nos han ayudado a entender ...

1.1 Análisis de Rendimiento

Primero que todo, a partir de las tablas de resultados, los oráculos extractivos nos marcan el límite superior del rendimiento que podríamos esperar de cualquier método extractivo. Cuanto valoremos los resultados de los NO oráculos, tomaremos estos valores como referencia para entender el margen de mejora que podríamos tener.

Por otro lado, el método Naive LEAD-3, nos sirve como una línea base simple para comparar los demás métodos y entender el dataset. Este nos dice cuenta información sobre los documentos se encuentra el las 3 primeras oraciones.

En un principio, bajo estar definiciones y teniendo en cuenta que todas las métricas parten de una referencia, un método extractivo NO podría superar el rendimiento de los oráculos extractivos. Los métodos podrían acercarse mucho a estos oráculos, pero no superarlos.

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 parecen ser métodos estables y robustos. Tiende a obtener buenos resultados en todos los aspectos, aunque no son los mejores en ninguno. Sobre todo, flaquean un poco en la fluidez. Algo normal en métodos extractivos porque solo juntan frases.

Por otro lado, en los modelos Sequence2Sequence podemos agrupar en Bart y Pegasus por un lado, y el Flan-t5 por otro. Bart y Pegasus parecen tener el mejor rendimiento en todas las métricas. Sin embargo, Flan-t5 parece tener un rendimiento en comparación inferior.... .

Esto algo esperado ya que Bart y Pegasus han sido finetuneados específicamente para la tarea de resumen automático EN ESTE DATASE, mientras que Flan-t5 es un modelo más generalista.

Los modelos causales (GPT2 y Llama3) parecen tener un rendimiento bastante inferior a los demás métodos. Esto puede ser debido a que estos modelos no están tan especializados en la tarea de resumen automático como los otros modelos abstractivos.

Viendo, podemos decir que el método extractivo más balanceado es TextRank, ya que obtiene resultados decentes en todos los aspectos sin destacar especialmente en ninguno.

Por otro lado, Bart diríamos que en el modelo abstractivo más balanceado. Está empatado con Pegasus en dos métricas, pero en las que Pegasus le supera, no lo hace por mucho.

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, en el cual no se ha entrenado ningún modelo específicamente, los resultados son un poco diferentes.

De primeras, es cierto que respecto a los métodos extractivos, el rendimiento es similar al del otro dataset. Sin embargo, en este caso, el método extractivo más balanceado sería LexRank, ya que obtiene mejores resultados que TextRank de forma consistente.

Ahora, si nos fijamos en los modelos abstractivos Sequence2Sequence, vemos que Flan-t5 obtiene los mejores resultados en casi todas las métricas. Esto es un cambio importante respecto al otro dataset, ya que Flan-t5 era el modelo con peor rendimiento.

El cambio de rendimiento de Flan-t5 puede ser debido a que este modelo es más generalista y no ha sido entrenado específicamente para el dataset CNN/DailyMail. Por lo tanto, en un dataset no visto, su rendimiento es mejor que el de los modelos más especializados como Bart y Pegasus.

Finalmente, los modelos causales (GPT2 y Llama3) siguen teniendo un rendimiento inferior al de los demás métodos, aunque en este caso, Llama3 obtiene mejores resultados que GPT2 en todas las métricas.

1.2 Análisis de Comparativo

1. ...

Primero que todo, elegiríamos un método abstractivo. Entre ellos, es complicado decidir entre 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 fácil de adaptar a distintos tipos de documentos y tareas en términos generales.

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

Teniendo esto presente, realiza un análsis en el que se contesten por lo menos a las siguientes preguntas:

1. Sí, negativo. El origen al que se le puede atribuir es que todas las métricas automáticas se basan en la comparación con los resúmenes de referencia. Por lo tanto, la abstractividad se verá penalizada. Si los resúmenes de referencia son muy extractivos, los métodos extractivos tendrán una ventaja al compararse con ellos, mientras que los métodos abstractivos podrían generar contenido que no esté presente en los resúmenes de referencia, lo que llevaría a una penalización en las métricas.

2. Sí, tienden a tener la misma relación negativa...