since 1995

Data Science Natural Language Processing NLP Papers Summary

Day 120: NLP Papers Summary – A Simple Theoretical Model Of Importance For Summarization

By Ryan 29th April 2020 No Comments

Objective and Contribution

Proposed a simple theoretical model to capture the information importance in summarisation. The model captures redundancy, relevance, and informativeness, all three of which contributes to the information importance in summarisation. We showcase how someone could use this framework to guide and improve summarisation systems. The contributions are as follows:

Loading [MathJax]/extensions/MathZoom.js arisation: redundancy, relevance, and informativeness

2. Formulate the Importance concept using the three key concepts in summarisation and show to interpret the results
3. Showed that our theoretical model of importance for summarisation has a good correlation with human summarisation, making it useful for guiding future empirical works

The Overall Framework

Semantic unit is considered a small piece of information. ω represents all the possible semantic units. A text input X is considered to be made up of many semantic units and so can be represented by a probability distribution $\mathbb{P}_X$ over Ω. $\mathbb{P}_X$ can simply means the frequency distribution of semantic units in the overall text. $\mathbb{P}_X(w_i)$ can be interpreted as the probability that the semantic unit w_i appears in text X or it could be interepreted as the contribution of w_i to the overall meaning of text X.

REDUNDANCY

The level of information presented in a summary is measured by entropy as follows:

$$H(S) = -\sum_{w_i} \mathbb{P}_S(w_i) x log(\mathbb{P}_S(w_i))$$

Entropy measures the coverage level and H(S) is maximised when every semantic unit in the summary only appears once and so the Redundancy formula is as follows:

$$Red(S) = H_{max} - H(S)$$

RELEVANCE

A relevant summary should be one that closely approximates the original text. In other words, a relevant summary should have the minimum loss of information. For us to measure relevancy, we would need to compare the probability distributions of the source document $\mathbb{P}_D$ and summary $\mathbb{P}_S$ using cross-entropy as follows:

$$Rel(S, D) = -CE(S, D) = \sum_{w_i} \mathbb{P}_S(w_i) x log(\mathbb{P}_D(w_i))$$

The formula is seen as the average surprise of producing S summary when expecting D source document. A summary S with low cross entropy (and so low surprise) implies low uncertainty about what were the original document. This is only possible if $\mathbb{P}_S$ is similar to $\mathbb{P}_D$.

KL divergence measures the loss of information when using source document D to generate summary S. The summary that minimises the KL divergence minimises redundancy and maximises relevance as it is the least biased (least redundant) summary matching D. The KL divergence connects redundancy and relevance as follows:

$$KL(S||D) = CE(S,D) - H(S)$$
$$-KL(S||D) = Rel(S,D) - Red(S)$$

INFORMATIVENESS

Informativeness introduce background knowledge K to capture the use of previous knowledge for summarisation. K is represented by $\mathbb{P}_K$ over all semantic units. The amount of new information in summary S is measured by the cross entropy between the summary and background knowledge as follows:

$$Inf(S,K) = CE(S,K)$$
$$Inf(S,K) = -\sum_{w_i} \mathbb{P}_S(w_i) x log(\mathbb{P}_K(w_i))$$

The cross entropy for relevance should be low as we want the summary to be as similar and relevant to the source document whereas the cross entropy for informativeness should be high as we are measuring the amount of background knowledge we used to generate the summary. This introduction of background knowledge allows us to customise the model depending on what kind of knowledge we want to include, whether that be domain-specific knowledge or user-specific knowledge or general knowledge. It also introduces the notion of update summarisation. Update summarisation involves summarising source document D having already seen document / summary U. Document / summary U could be modelled by background knowledge K, which makes U a previous knowledge.

IMPORTANCE

Importance is the metric that guides what information should be included in the summary. Given a user with knowledge K, the summary should be generated with the objective to bring the most new information to the user. Therefore, for each semantic unit, we need a function $f(d_i, k_i)$ that takes in the probability of semantic unit in source document D ($d_i = \mathbb{P}_I$ and

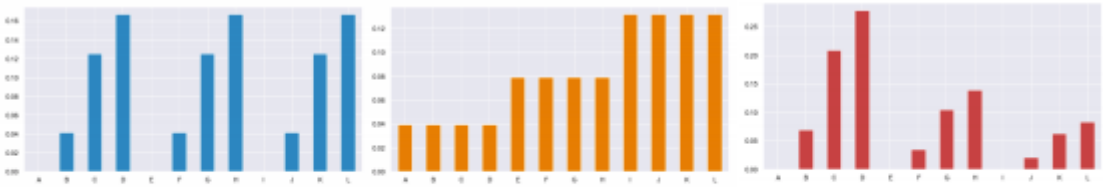
background knowledge ($k_i = \mathbb{P}_K(w_i)$), to determine its importance. The function $f(d_i, k_i)$ as four requirements:

1. *Informativeness*. If two semantic units are equally important in the source document, we would prefer the one that are more informative, which it's governed by background knowledge
2. *Relevance*. If two semantic units are equally informative, then we would prefer the semantic unit that's more important in the source document
3. *Additivity*. This is a consistency constraint to allow for addition of information measures
4. *Normalisation*. To ensure that the funtion is a valid distribution

SUMMARY SCORING FUNCTION

$\mathbb{P}_{(\frac{D}{K})}$ encodes the relative importance of semantic units, the trade-off between relevance and informativeness. An example of what this distribution would capture is that if the semantic unit is important in source document but it's not known in background knowledge, then $\mathbb{P}_{(\frac{D}{K})}$ is very high for that semantic unit as it is very desirable to be included in the summary as it increases the knowledge gap. This is illustrated in the figure below. The summary should be non-redundant and best approximate $\mathbb{P}_{(\frac{D}{K})}$ as follows:

$$S* = argmax\theta_I = argminKL(S||\mathbb{P}_{(\frac{D}{K})})$$
$$\theta_I(S, D, K) = -KL(\mathbb{S}||\mathbb{P}_{(\frac{D}{K})})$$



(a) ditribution $\mathbb{P}_D$ (b) distribution $\mathbb{P}_K$ (c) distribution $\mathbb{P}_{\frac{D}{K}}$

Figure 1: figure 1a represents an example distribution of sources, figure 1b an example distribution of background knowledge and figure 1c is the resulting target distribution that summaries should approximate.

SUMMARISABILITY

We can use the $\mathbb{P}_{(\frac{D}{K})}$ to measure how many good summaries can be extracted from the distribution as follows:

Loading [MathJax]/extensions/MathZoom.js

$$H_{\frac{D}{K}} = H(\mathbb{P}_{(\frac{D}{K})})$$

If $H_{\frac{D}{K}}$ is high, then there are many similar good summaries that can be generated from the distribution. Conversely, if it's low, there are only few good summaries. In terms of the summary scoring function, another way of expressing it is as follows:

$$\theta_I(S, D, K) = -Red(S) + \alpha Rel(S, D) + \beta Inf(S, K)$$

Maximising θ_I is equivalent of maximising the relevance and informativeness while minimising the redundancy, which it's exactly what we want in a high quality summary. α represents the strength of the Relevance component and β represents the strength of the Informativeness component. This means that H(S), CE(S, D), and CE(S, K) are three independent factors that affects the Importance concept.

POTENTIAL INFORMATION

So far, we have connected summary S with source document D using relevance and summary S with background knowledge K using informativeness. However, we could also connect source document D with background knowledge K. We can extract a lot of new information from source document D if it strongly differs from K. The computation of this is the same as Informativeness except it is between source document D and background knowledge K. This new cross-entropy represents the maximum information gain that's possible from source document D given background knowledge K.

Experiments

We used two evaluation datasets: TAC-2008 and TAC-2009. The datasets focus on two different summarisation tasks: normal and update summarisation for multi-document. Background knowledge K, α, and β are the parameters of our theoretical model for summarisation. We have set $\alpha = \beta = 1$ and the background knowledge K to either be frequency distribution over words in background documents or probability distribution over all words from source documents.

CORRELATION WITH HUMAN JUDGEMENTS

We assess how well our quantities correlate with human judgements. Each quan f our fr nces for summary and so we can evaluate how well they

correlate with human judgement. The results are showcase below. Out of the three quanti ^ , it seems that relevance has the highest correlation with human judgements. The inclusion of background knowledge works better with update summarisation as expected. Lastly, the θ_I gives the best performance in both types of summarisation. Individual quantity did not have strong performance on their own but once they are put together, it gives us a reliable strong summary scoring function.

COMPARISON WITH REFERENCE SUMMARIES

Ideally we would want our generated summaries (using $\mathbb{P}_{(\frac{D}{K})}$) to be similar to human reference summaries ($\mathbb{P}_R$). We scored both summaries using θ_I and found that human reference summaries scored significantly higher than our generated summaries, proving the reliability of our scoring function.

Conclusion and Future Work

Importance unifies the three common metrics of redundancy, relevance, and informativeness when it comes to summarisation and tells us which information to discard or include in the final summary. Background knowledge and semantic units choice are open parameters of the theoretical model, which means that they are open for experimentation / exploration. N-grams are good approximation of semantic units but what other granularity could we consider here?

Potential future work for background knowledge could be to use the framework to learn knowledge from the data. Specifically, you can train a model to learn background knowledge such that the model has the highest correlation with human judgements. If you aggregate all the information over all the users and topics, you can find the generic background knowledge. If you aggregate all the users but in one particular topic, you can find topic-specific background knowledge and similar work can be done for a single user.

Source: https://www.aclweb.org/anthology/P19-1101.pdf

Ryan

Data Scientist

Previous Post

Day 119: NLP Papers Summary - An Argument-Annotated Corpus of Scientific Publications

Next Post

Day 121: NLP Papers Summary - Concept Pointer Network for Abstractive Summarization

Data Science Natural Language Processing NLP Papers Summary

Day 365: NLP Papers Summary – A Survey on Knowledge Graph Embedding

Ryan
30th December 2020

Loading [MathJax]/extensions/MathZoom.js

Data Science Implementation Natural Language Processing

Day 364: Ryan's PhD Journey – OpenKE-PyTorch Library Analysis + code snippets for 11 KE models

Ryan
29th December 2020

Loading [MathJax]/extensions/MathZoom.js