A Simple Theoretical Model of Importance for Summarization

Maxime Peyrard*
EPFL
maxime.peyrard@epfl.ch

Abstract

Research on summarization has mainly been driven by empirical approaches, crafting systems to perform well on standard datasets with the notion of information *Importance* remaining latent. We argue that establishing theoretical models of *Importance* will advance our understanding of the task and help to further improve summarization systems. To this end, we propose simple but rigorous definitions of several concepts that were previously used only intuitively in summarization: *Redundancy*, *Relevance*, and *Informativeness*. Importance arises as a single quantity naturally unifying these concepts. Additionally, we provide intuitions to interpret the proposed quantities and experiments to demonstrate the potential of the framework to inform and guide subsequent works.

1 Introduction

Summarization is the process of identifying the most *important information* from a source to produce a comprehensive output for a particular user and task (Mani, 1999). While producing readable outputs is a problem shared with the field of *Natural Language Generation*, the core challenge of summarization is the identification and selection of *important information*. The task definition is rather intuitive but involves vague and undefined terms such as *Importance* and *Information*.

Since the seminal work of Luhn (1958), automatic text summarization research has focused on empirical developments, crafting summarization systems to perform well on standard datasets leaving the formal definitions of *Importance* latent (Das and Martins, 2010; Nenkova and McKeown, 2012). This view entails collecting datasets, defining evaluation metrics and iteratively selecting the best-performing systems either via supervised learning or via repeated comparison of unsupervised systems (Yao et al., 2017).

Such solely empirical approaches may lack guidance as they are often not motivated by more general theoretical frameworks. While these approaches have facilitated the development of practical solutions, they only identify signals correlating with the vague human intuition of *Importance*. For instance, structural features like centrality and repetitions are still among the most used proxies for *Importance* (Yao et al., 2017; Kedzie et al., 2018). However, such features just correlate with *Importance* in standard datasets. Unsurprisingly, simple adversarial attacks reveal their weaknesses (Zopf et al., 2016).

We postulate that theoretical models of *Importance* are beneficial to organize research and guide future empirical works. Hence, in this work, we propose a simple definition of information importance within an abstract theoretical framework. This requires the notion of information, which has received a lot of attention since the work from Shannon (1948) in the context of communication theory. Information theory provides the means to rigorously discuss the abstract concept of information, which seems particularly well suited as an entry point for a theory of summarization. However, information theory concentrates on uncertainty (entropy) about which message was chosen from a set of possible messages, ignoring the semantics of messages (Shannon, 1948). Yet, summarization is a lossy semantic compression depending on background knowledge.

In order to apply information theory to summarization, we assume texts are represented by probability distributions over so-called *semantic units* (Bao et al., 2011). This view is compatible with the common distributional embedding representation of texts rendering the presented framework applicable in practice. When applied

Research partly done at UKP Lab from TU Darmstadt.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1059–1073
Florence, Italy, July 28 - August 2, 2019.

to semantic symbols, the tools of information theory indirectly operate at the semantic level (Carnap and Bar-Hillel, 1953; Zhong, 2017).

Contributions:

- We define several concepts intuitively connected to summarization: *Redundancy*, *Relevance* and *Informativeness*. These concepts have been used extensively in previous summarization works and we discuss along the way how our framework generalizes them.
- From these definitions, we formulate properties required from a useful notion of *Importance* as the quantity unifying these concepts. We provide intuitions to interpret the proposed quantities.
- Experiments show that, even under simplifying assumptions, these quantities correlates well with human judgments making the framework promising in order to guide future empirical works.

2 Framework

2.1 Terminology and Assumptions

We call *semantic unit* an atomic piece of information (Zhong, 2017; Cruse, 1986). We note Ω the set of all possible semantic units.

A text X is considered as a semantic source emitting semantic units as envisioned by Weaver (1953) and discussed by Bao et al. (2011). Hence, we assume that X can be represented by a probability distribution $\mathbb{P}_X$ over the semantic units Ω.

Possible interpretations:
One can interpret $\mathbb{P}_X$ as the frequency distribution of semantic units in the text. Alternatively, $\mathbb{P}_X(\omega_i)$ can be seen as the (normalized) likelihood that a text X entails an atomic information ω_i (Carnap and Bar-Hillel, 1953). Another interpretation is to view $\mathbb{P}_X(\omega_i)$ as the normalized contribution (utility) of ω_i to the overall meaning of X (Zhong, 2017).

Motivation for semantic units:
In general, existing semantic information theories either postulate or imply the existence of semantic units (Carnap and Bar-Hillel, 1953; Bao et al., 2011; Zhong, 2017). For example, the *Theory of Strongly Semantic Information* produced by Floridi (2009) implies the existence of semantic units (called information units in his work). Building on this, Tsvetkov (2014) argued that the original theory of Shannon can operate at the semantic level by relying on semantic units.

In particular, existing semantic information theories imply the existence of semantic units in formal semantics (Carnap and Bar-Hillel, 1953), which treat natural languages as formal languages (Montague, 1970). In general, lexical semantics (Cruse, 1986) also postulates the existence of elementary constituents called minimal semantic constituents. For instance, with frame semantics (Fillmore, 1976), frames can act as semantic units.

Recently, distributional semantics approaches have received a lot of attention (Turian et al., 2010; Mikolov et al., 2013b). They are based on the distributional hypothesis (Harris, 1954) and the assumption that meaning can be encoded in a vector space (Turney and Pantel, 2010; Erk, 2010). These approaches also search latent and independent components that underlie the behavior of words (Gábor et al., 2017; Mikolov et al., 2013a).

While different approaches to semantics postulate different basic units and different properties for them, they have in common that *meaning arises from a set of independent and discrete units*. Thus, the semantic units assumption is general and has minimal commitment to the actual nature of semantics. This makes the framework compatible with most existing semantic representation approaches. Each approach specifies these units and can be plugged in the framework, e.g., frame semantics would define units as frames, topic models (Allahyari et al., 2017) would define units as topics and distributional representations would define units as dimensions of a vector space.

In the following paragraphs, we represent the source document(s) D and a candidate summary S by their respective distributions $\mathbb{P}_D$ and $\mathbb{P}_S$.[1]

2.2 Redundancy

Intuitively, a summary should contain a lot of information. In information-theoretic terms, the *amount of information* is measured by Shannon's

[1] We sometimes note X instead of $\mathbb{P}_X$ when it is not ambiguous

entropy. For a summary S represented by $\mathbb{P}_S$:

$$H(S) = -\sum_{\omega_i} \mathbb{P}_S(\omega_i) \cdot \log(\mathbb{P}_S(\omega_i)) \quad (1)$$

$H(S)$ is maximized for a uniform probability distribution when every semantic unit is present only once in S: $\forall(i,j), \mathbb{P}_S(\omega_i) = \mathbb{P}_S(\omega_j)$. Therefore, we define *Redundancy*, our first quantity relevant to summarization, via entropy:

$$Red(S) = H_{max} - H(S) \quad (2)$$

Since $H_{max} = \log|\Omega|$ is a constant indepedent of S, we can simply write: $Red(S) = -H(S)$.

Redundancy in Previous Works:
By definition, entropy encompasses the notion of maximum coverage. Low redundancy via maximum coverage is the main idea behind the use of submodularity (Lin and Bilmes, 2011). Submodular functions are generalizations of coverage functions which can be optimized greedily with guarantees that the result would not be far from optimal (Fujishige, 2005). Thus, they have been used extensively in summarization (Sipos et al., 2012; Yogatama et al., 2015). Otherwise, low redundancy is usually enforced during the extraction/generation procedures like MMR (Carbonell and Goldstein, 1998).

2.3 Relevance

Intuitively, observing a summary should reduce our uncertainty about the original text. A summary approximates the original source(s) and this approximation should incur a minimum loss of information. This property is usually called *Relevance*.

Here, estimating *Relevance* boils down to comparing the distributions $\mathbb{P}_S$ and $\mathbb{P}_D$, which is done via the cross-entropy $Rel(S,D) = -CE(S,D)$:

$$Rel(S,D) = \sum_{\omega_i} \mathbb{P}_S(\omega_i) \cdot \log(\mathbb{P}_D(\omega_i)) \quad (3)$$

The cross-entropy is interpreted as the average surprise of observing S while expecting D. A summary with a low expected surprise produces a low uncertainty about what were the original sources. This is achieved by exhibiting a distribution of semantic units similar to the one of the source documents: $\mathbb{P}_S \approx \mathbb{P}_D$.

Furthermore, we observe the following connection with *Redundancy*:

$$\begin{aligned} KL(S||D) &= CE(S,D) - H(S) \\ -KL(S||D) &= Rel(S,D) - Red(S) \end{aligned} \quad (4)$$

KL divergence is the information loss incurred by using D as an approximation of S (i.e., the uncertainty about D arising from observing S instead of D). A summarizer that minimizes the KL divergence minimizes *Redundancy* while maximizing *Relevance*.

In fact, this is an instance of the *Kullback Minimum Description Principle* (MDI) (Kullback and Leibler, 1951), a generalization of the *Maximum Entropy Principle* (Jaynes, 1957): the summary minimizing the KL divergence is the least biased (i.e., least redundant or with highest entropy) summary matching D. In other words, this summary fits D while inducing a minimum amount of *new* information. Indeed, any *new* information is necessarily biased since it does not arise from observations in the sources. The MDI principle and KL divergence unify *Redundancy* and *Relevance*.

Relevance in Previous Works:
Relevance is the most heavily studied aspect of summarization. In fact, by design, most unsupervised systems model *Relevance*. Usually, they used the idea of *topical frequency* where the most frequent topics from the sources must be extracted. Then, different notions of *topics* and counting heuristics have been proposed. We briefly discuss these developments here.

Luhn (1958) introduced the simple but influential idea that sentences containing the most important words are most likely to embody the original document. Later, Nenkova et al. (2006) showed experimentally that humans tend to use words appearing frequently in the sources to produce their summaries. Then, Vanderwende et al. (2007) developed the system *SumBasic*, which scores each sentence by the average probability of its words.

The same ideas can be generalized to n-grams. A prominent example is the ICSI system (Gillick and Favre, 2009) which extracts frequent bigrams. Despite being rather simple, ICSI produces strong and still close to state-of-the-art summaries (Hong et al., 2014).

Different but similar words may refer to the same topic and should not be counted separately.

This observation gave rise to a set of important techniques based on topic models (Allahyari et al., 2017). These approaches cover sentence clustering (McKeown et al., 1999; Radev et al., 2000; Zhang et al., 2015), lexical chains (Barzilay and Elhadad, 1999), Latent Semantic Analysis (Deerwester et al., 1990) or Latent Dirichlet Allocation (Blei et al., 2003) adapted to summarization (Hachey et al., 2006; Daumé III and Marcu, 2006; Wang et al., 2009; Davis et al., 2012). Approaches like hLDA can exploit repetitions both at the word and at the sentence level (Celikyilmaz and Hakkani-Tur, 2010).

Graph-based methods form another particularly powerful class of techniques to estimate the frequency of topics, e.g., via the notion of centrality (Mani and Bloedorn, 1997; Mihalcea and Tarau, 2004; Erkan and Radev, 2004). A significant body of research was dedicated to tweak and improve various components of graph-based approaches. For example, one can investigate different similarity measures (Chali and Joty, 2008). Also, different weighting schemes between sentences have been investigated (Leskovec et al., 2005; Wan and Yang, 2006).

Therefore, in existing approaches, the topics (i.e., atomic units) were words, n-grams, sentences or combinations of these. The general idea of preferring *frequent topics* based on various counting heuristics is formalized by cross-entropy. Indeed, requiring the summary to minimize the cross-entropy with the source documents implies that frequent topics in the sources should be extracted first.

An interesting line of work is based on the assumption that the best sentences are the ones that permit the best reconstruction of the input documents (He et al., 2012). It was refined by a stream of works using distributional similarities (Li et al., 2015; Liu et al., 2015; Ma et al., 2016). There, the atomic units are the dimensions of the vector spaces. This information bottleneck idea is also neatly captured by the notion of cross-entropy which is a measure of information loss. Alternatively, (Daumé and Marcu, 2002) viewed summarization as a noisy communication channel which is also rooted in information theory ideas. (Wilson and Sperber, 2008) provide a more general and less formal discussion of relevance in the context of Relevance Theory (Lavrenko, 2008).

2.4 Informativeness

Relevance still ignores other potential sources of information such as previous knowledge or preconceptions. We need to further extend the contextual boundary. Intuitively, a summary is informative if it induces, for a user, a great change in her knowledge about the world. Therefore, we introduce K, the background knowledge (or preconceptions about the task). K is represented by a probability distribution $\mathbb{P}_K$ over semantic units Ω.

Formally, the amount of *new* information contained in a summary S is given by the cross-entropy $Inf(S,K) = CE(S,K)$:

$$Inf(S,K) = -\sum_{\omega_i} \mathbb{P}_S(\omega_i) \cdot \log(\mathbb{P}_K(\omega_i)) \quad (5)$$

For *Relevance* the cross-entropy between S and D should be low. However, for *Informativeness*, the cross-entropy between S and K should be high because we measure the amount of new information induced by the summary in our knowledge.

Background knowledge is modeled by assigning a high probability to known semantic units. These probabilities correspond to the strength of ω_i in the user's memory. A simple model could be the uniform distribution over known information: $\mathbb{P}_K(\omega_i)$ is $\frac{1}{n}$ if the user knows ω_i, and 0 otherwise. However, K can control other variants of the summarization task: A personalized K_p models the preferences of a user by setting low probabilities to the semantic units of interest. Similarly, a query Q can be encoded by setting low probability to semantic units related to Q. Finally, there is a natural formulation of update summarization. Let U and D be two sets of documents. Update summarization consists in summarizing D given that the user has already seen U. This is modeled by setting $K = U$, considering U as previous knowledge.

Informativeness in Previous Works:
The modelization of *Informativeness* has received less attention by the summarization community. The problem of identifying stopwords originally faced by Luhn (1958) could be addressed by developments in the field of information retrieval using background corpora like TF·IDF (Sparck Jones, 1972). Based on the same intuition, Dunning (1993) outlined an alternative way of identifying highly descriptive words: the *log-likelihood ratio* test. Words identified with such

techniques are known to be useful in news summarization (Harabagiu and Lacatusu, 2005).

Furthermore, Conroy et al. (2006) proposed to model background knowledge by a large random set of news articles. In update summarization, Delort and Alfonseca (2012) used Bayesian topic models to ensure the extraction of informative summaries. Louis (2014) investigated background knowledge for update summarization with Bayesian surprise. This is comparable to the combination of *Informativeness* and *Redundancy* in our framework when semantic units are n-grams. Thus, previous approaches to *Informativeness* generally craft an alternate background distribution to model the *a-priori* importance of units. Then, units from the document rare in the background are preferred, which is captured by maximizing the cross-entropy between the summary and K. Indeed, unfrequent units in the background would be preferred in the summary because they would be surprising (i.e., informative) to an average user.

2.5 Importance

Since *Importance* is a measure that guides which choices to make when discarding semantic units, we must devise a way to encode their relative importance. Here, this means finding a probability distribution unifying D and K by encoding expectations about which semantic units should appear in a summary.

Informativeness requires a biased summary (w.r.t. K) and *Relevance* requires an unbiased summary (w.r.t. D). Thus, a summary should, by using only information available in D, produce what brings the most new information to a user with knowledge K. This could formalize a common intuition in summarization that units frequent in the source(s) but rare in the background are important.

Formally, let $d_i = \mathbb{P}_D(\omega_i)$ be the probability of the unit ω_i in the source D. Similarly, we note $k_i = \mathbb{P}_K(\omega_i)$. We seek a function $f(d_i, k_i)$ encoding the importance of unit ω_i. We formulate simple requirements that f should satisfy:

- Informativeness: $\forall i \neq j$, if $d_i = d_j$ and $k_i > k_j$ then $f(d_i, k_i) < f(d_j, k_j)$
- Relevance: $\forall i \neq j$, if $d_i > d_j$ and $k_i = k_j$ then $f(d_i, k_i) > f(d_j, k_j)$
- Additivity: $I(f(d_i, k_i)) \equiv \alpha I(d_i) + \beta I(k_i)$ (I is the information measure from Shannon's theory (Shannon, 1948))
- Normalization: $\sum_i f(d_i, k_i) = 1$

The first requirement states that, for two semantic units equally represented in the sources, we prefer the more informative one. The second requirement is an analogous statement for *Relevance*. The third requirement is a consistency constraint to preserve additivity of the information measures (Shannon, 1948). The fourth requirement ensures that f is a valid distribution.

Theorem 1. *The functions satisfying the previous requirements are of the form:*

$$\mathbb{P}_{\frac{D}{K}}(\omega_i) = \frac{1}{C} \cdot \frac{d_i^\alpha}{k_i^\beta} \tag{6}$$

$$C = \sum_i \frac{d_i^\alpha}{k_i^\beta}, \ \alpha, \beta \in \mathbb{R}^+ \tag{7}$$

C is the normalizing constant. The parameters α and β represent the strength given to *Relevance* and *Informativeness* respectively which is made clearer by equation (11). The proof is provided in appendix B.

Summary scoring function:
By construction, a candidate summary should approximate $\mathbb{P}_{\frac{D}{K}}$, which encodes the relative importance of semantic units. Furthermore, the summary should be non-redundant (i.e., high entropy). These two requirements are unified by the Kullback MDI principle: The least biased summary S^* that best approximates the distribution $\mathbb{P}_{\frac{D}{K}}$ is the solution of:

$$S^* = \underset{S}{\operatorname{argmax}}\, \theta_I = \underset{S}{\operatorname{argmin}}\, KL(S||\mathbb{P}_{\frac{D}{K}}) \tag{8}$$

Thus, we note θ_I as the quantity that scores summaries:

$$\theta_I(S, D, K) = -KL(\mathbb{P}_S, ||\mathbb{P}_{\frac{D}{K}}) \tag{9}$$

Interpretation of $\mathbb{P}_{\frac{D}{K}}$:
$\mathbb{P}_{\frac{D}{K}}$ can be viewed as an *importance-encoding distribution* because it encodes the relative importance of semantic units and gives an overall target for the summary.

For example, if a semantic unit ω_i is prominent in D ($\mathbb{P}_D(\omega_i)$ is high) and not known in K ($\mathbb{P}_D(\omega_i)$ is low), then $\mathbb{P}_{\frac{D}{K}}(\omega_i)$ is very high,

which means very desired in the summary. Indeed, choosing this unit will fill the gap in the knowledge K while matching the sources.

Figure 1 illustrates how this distribution behaves with respect to D and K (for $\alpha = \beta = 1$).

Summarizability:
The target distribution $\mathbb{P}_{\frac{D}{K}}$ may exhibit different properties. For example, it might be clear which semantic units should be extracted (i.e., a spiky probability distribution) or it might be unclear (i.e., many units have more or less the same importance score). This can be quantified by the entropy of the importance-encoding distribution:

$$H_{\frac{D}{K}} = H(\mathbb{P}_{\frac{D}{K}}) \quad (10)$$

Intuitively, this measures the number of possibly good summaries. If $H_{\frac{D}{K}}$ is low then $\mathbb{P}_{\frac{D}{S}}$ is spiky and there is little uncertainty about which semantic units to extract (few possible *good* summaries). Conversely, if the entropy is high, many equivalently *good* summaries are possible.

Interpretation of θ_I:
To better understand θ_I, we remark that it can be expressed in terms of the previously defined quantities:

$$\theta_I(S, D, K) \equiv -Red(S) + \alpha Rel(S, D) \quad (11)$$
$$+ \beta Inf(S, K) \quad (12)$$

Equality holds up to a constant term $\log C$ independent from S. Maximizing θ_I is equivalent to maximizing *Relevance* and *Informativeness* while minimizing *Redundancy*. Their relative strength are encoded by α and β.

Finally, $H(S)$, $CE(S, D)$ and $CE(S, K)$ are the three independent components of *Importance*.

It is worth noting that each previously defined quantity: Red, Rel and Inf are measured in bits (using base 2 for the logarithm). Then, θ_I is also an information measure expressed in bits. Shannon (1948) initially axiomatized that information quantities should be additive and therefore θ_I arising as the sum of other information quantities is unsurprising. Moreover, we ensured additivity with the third requirement of $\mathbb{P}_{\frac{D}{K}}$.

2.6 Potential Information

Relevance relates S and D, *Informativeness* relates S and K, but we can also connect D and K. Intuitively, we can extract a lot of new information from D only when K and D are different.

With the same argument laid out for *Informativeness*, we can define the amount of potential information as the average surprise of observing D while already knowing K. Again, this is given by the cross-entropy $PI_K(D) = CE(D, K)$:

$$PI_K(D) = -\sum_{\omega_i} \mathbb{P}_D(\omega_i) \cdot \log(\mathbb{P}_K(\omega_i)) \quad (13)$$

Previously, we stated that a summary should aim, using only information from D, to offer the maximum amount of new information with respect to K. $PI_K(D)$ can be understood as *Potential Information* or maximum *Informativeness*, the maximum amount of new information that a summary can extract from D while knowing K. A summary S cannot extract more than $PI_K(D)$ bits of information (if using only information from D).

3 Experiments

3.1 Experimental setup

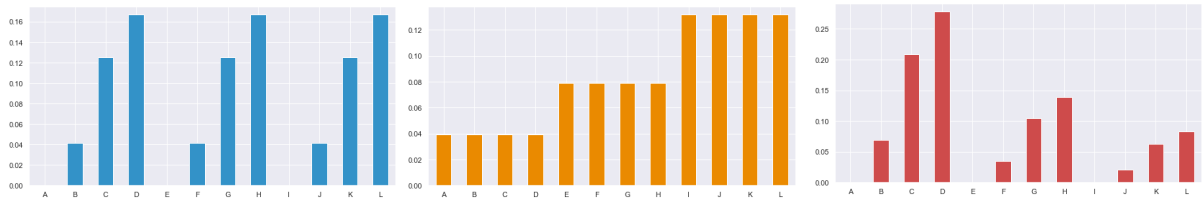
To further illustrate the workings of the formula, we provide examples of experiments done with a simplistic choice for semantic units: words. Even with simple assumptions θ_I is a meaningful quantity which correlates well with human judgments.

Data:
We experiment with standard datasets for two different summarization tasks: generic and update multi-document summarization.

We use two datasets from the Text Analysis Conference (TAC) shared task: TAC-2008 and TAC-2009.[2] In the update part, 10 new documents (B documents) are to be summarized assuming that the first 10 documents (A documents) have already been seen. The generic task consists in summarizing the initial document set (A).

For each topic, there are 4 human reference summaries and a manually created Pyramid set (Nenkova et al., 2007). In both editions, all system summaries and the 4 reference summaries were manually evaluated by NIST assessors for readability, content selection (with Pyramid) and overall responsiveness. At the time of the shared tasks, 57 systems were submitted to TAC-2008 and 55 to TAC-2009.

[2] http://tac.nist.gov/2009/Summarization/, http://tac.nist.gov/2008/



(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.

Setup and Assumptions:
To keep the experiments simple and focused on illustrating the formulas, we make several simplistic assumptions. First, we choose words as semantic units and therefore texts are represented as frequency distributions over words. This assumption was already employed by previous works using information-theoretic tools for summarization (Haghighi and Vanderwende, 2009). While it is limiting, this remains a simple approximation letting us observe the quantities in action.

K, α and β are the parameters of the theory and their choice is subject to empirical investigation. Here, we make simple choices: for update summarization, K is the frequency distribution over words in the background documents (A). For generic summarization, K is the uniform probability distribution over all words from the source documents. Furthermore, we use $\alpha = \beta = 1$.

3.2 Correlation with humans

First, we measure how well the different quantities correlate with human judgments. We compute the score of each system summary according to each quantity defined in the previous section: Red, Rel, Inf, $\theta_I(S, D, K)$. We then compute the correlations between these scores and the manual Pyramid scores. Indeed, each quantity is a summary scoring function and could, therefore, be evaluated based on its ability to correlate with human judgments (Lin and Hovy, 2003). Thus, we also report the performances of the summary scoring functions from several standard baselines: **Edmundson** (Edmundson, 1969) which scores sentences based on 4 methods: term frequency, presence of cue-words, overlap with title and position of the sentence. **LexRank** (Erkan and Radev, 2004) is a popular graph-based approach which scores sentences based on their centrality in a sentence similarity graph. **ICSI** (Gillick and Favre, 2009) extracts a summary by solving a maximum coverage problem considering the most frequent bigrams in the source documents. **KL** and **JS** (Haghighi and Vanderwende, 2009) which measure the divergence between the distribution of words in the summary and in the sources. Furthermore, we report two baselines from Louis (2014) which account for background knowledge: KL_{back} and JS_{back} which measure the divergence between the distribution of the summary and the background knowledge K. Further details concerning baseline scoring functions can be found in appendix A.

We measure the correlations with Kendall's τ, a rank correlation metric which compares the orders induced by both scored lists. We report results for both generic and update summarization averaged over all topics for both datasets in table 1.

In general, the modelizations of *Relevance* (based only on the sources) correlate better with human judgments than other quantities. Metrics accounting for background knowledge work better in the update scenario. This is not surprising as the background knowledge K is more meaningful in this case (using the previous document set).

We observe that JS divergence gives slightly better results than KL. Even though KL is more theoretically appealing, JS is smoother and usually works better in practice when distributions have different supports (Louis and Nenkova, 2013).

Finally, θ_I significantly[3] outperforms all baselines in both the generic and the update case. *Red*, *Rel* and *Inf* are not particularly strong on their own, but combined together they yield a strong summary scoring function θ_I. Indeed, each quantity models only one aspect of content selection, only together they form a strong signal for *Importance*.

[3] at 0.01 with significance testing done with a t-test to compare two means

We need to be careful when interpreting these results because we made several strong assumptions: by choosing n-grams as semantic units and by choosing K rather arbitrarily. Nevertheless, these are promising results. By investigating better text representations and more realistic K, we should expect even higher correlations.

We provide a qualitative example on one topic in appendix C with a visualization of $\mathbb{P}_{\frac{D}{K}}$ in comparison to reference summaries.

	Generic	Update
ICSI	.178	.139
Edm.	.215	.205
LexRank	.201	.164
KL	.204	.176
JS	.225	.189
KL_{back}	.110	.167
JS_{back}	.066	.187
Red	.098	.096
Rel	.212	.192
Inf	.091	.086
θ_I	**.294**	**.211**

Table 1: Correlation of various information-theoretic quantities with human judgments measured by Kendall's τ on generic and update summarization.

3.3 Comparison with Reference Summaries

Intuitively, the distribution $\mathbb{P}_{\frac{D}{K}}$ should be similar to the probability distribution $\mathbb{P}_R$ of the human-written reference summaries.

To verify this, we scored the system summaries and the reference summaries with θ_I and checked whether there is a significant difference between the two lists.[4] We found that θ_I scores reference summaries significantly higher than system summaries. The $p-$value, for the generic case, is $9.2\text{e}{-6}$ and $1.1\text{e}{-3}$ for the update case. Both are much smaller than the $1\text{e}{-2}$ significance level. Therefore, θ_I is capable of distinguishing systems summaries from human written ones. For comparison, the best baseline (JS) has the following $p-$values: $8.2\text{e}{-3}$ (Generic) and $4.5\text{e}{-2}$ (Update). It does not pass the $1\text{e}{-2}$ significance level for the update scenario.

[4] with standard t-test for comparing two related means.

4 Conclusion and Future Work

In this work, we argued for the development of theoretical models of *Importance* and proposed one such framework. Thus, we investigated a theoretical formulation of the notion of *Importance*. In a framework rooted in information theory, we formalized several summary-related quantities like: *Redundancy*, *Relevance* and *Informativeness*. *Importance* arises as the notion unifying these concepts. More generally, *Importance* is the measure that guides which choices to make when information must be discarded. The introduced quantities generalize the intuitions that have previously been used in summarization research.

Conceptually, it is straightforward to build a system out of θ_I once a semantic units representation and a K have been chosen. A summarizer intends to extract or generate a summary maximizing θ_I. This fits within the general optimization framework for summarization (McDonald, 2007; Peyrard and Eckle-Kohler, 2017b; Peyrard and Gurevych, 2018)

The background knowledge and the choice of semantic units are free parameters of the theory. They are design choices which can be explored empirically by subsequent works. Our experiments already hint that strong summarizers can be developed from this framework. Characters, character n-grams, morphemes, words, n-grams, phrases, and sentences do not actually qualify as semantic units. Even though previous works who relied on information theoretic motivation (Lin et al., 2006; Haghighi and Vanderwende, 2009; Louis and Nenkova, 2013; Peyrard and Eckle-Kohler, 2016) used some of them as support for probability distributions, they are neither atomic nor independent. It is mainly because they are surface forms whereas semantic units are abstract and operate at the semantic level. However, they might serve as convenient approximations. Then, interesting research questions arise like *Which granularity offers a good approximation of semantic units? Can we automatically learn good approximations?* N-grams are known to be useful, but other granularities have rarely been considered together with information-theoretic tools.

For the background knowledge K, a promising direction would be to use the framework to actually learn it from data. In particular, one can apply supervised techniques to automatically search for K, α and β: finding the values of these parame-

ters such that θ_I has the best correlation with human judgments. By aggregating over many users and many topics one can find a generic K: what, on average, people consider as known when summarizing a document. By aggregating over different people but in one domain, one can uncover a domain-specific K. Similarly, by aggregating over many topics for one person, one would find a personalized K.

These consistute promising research directions for future works.

Acknowledgements

This work was partly supported by the German Research Foundation (DFG) as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1, and via the German-Israeli Project Cooperation (DIP, grant No. GU 798/17-1). We also thank the anonymous reviewers for their comments.

References

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text Summarization Techniques: A Brief Survey. *International Journal of Advanced Computer Science and Applications*, 8(10).

Jie Bao, Prithwish Basu, Mike Dean, Craig Partridge, Ananthram Swami, Will Leland, and James A Hendler. 2011. Towards a theory of semantic communication. In *Network Science Workshop (NSW), 2011 IEEE*, pages 110–117. IEEE.

Regina Barzilay and Michael Elhadad. 1999. Using Lexical Chains for Text Summarization. *Advances in Automatic Text Summarization*, pages 111–121.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336.

Rudolf Carnap and Yehoshua Bar-Hillel. 1953. An Outline of a Theory of Semantic Information. *British Journal for the Philosophy of Science.*, 4.

Asli Celikyilmaz and Dilek Hakkani-Tur. 2010. A Hybrid Hierarchical Model for Multi-Document Summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 815–824, Uppsala, Sweden. Association for Computational Linguistics.

Yllias Chali and Shafiq R. Joty. 2008. Improving the performance of the random walk model for answering complex questions. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 9–12. Association for Computational Linguistics.

John M. Conroy, Judith D. Schlesinger, and Dianne P. O'Leary. 2006. Topic-Focused Multi-Document Summarization Using an Approximate Oracle Score. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 152–159, Sydney, Australia. Association for Computational Linguistics.

D.A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge, UK.

Dipanjan Das and André F. T. Martins. 2010. A Survey on Automatic Text Summarization. *Literature Survey for the Language and Statistics II Course at CMU*.

Hal Daumé, III and Daniel Marcu. 2002. A Noisy-channel Model for Document Compression. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 449–456.

Hal Daumé III and Daniel Marcu. 2006. Bayesian Query-Focused Summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 305–312. Association for Computational Linguistics.

Sashka T. Davis, John M. Conroy, and Judith D. Schlesinger. 2012. OCCAMS–An Optimal Combinatorial Covering Algorithm for Multi-document Summarization. In *Proceeding of the 12th International Conference on Data Mining Workshops (ICDMW)*, pages 454–463. IEEE.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Jean-Yves Delort and Enrique Alfonseca. 2012. DualSum: A Topic-model Based Approach for Update Summarization. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 214–223.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational linguistics*, 19(1):61–74.

H. P. Edmundson. 1969. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285.

Katrin Erk. 2010. What is Word Meaning, Really? (and How Can Distributional Models Help Us Describe It?). In *Proceedings of the 2010 workshop on geometrical models of natural language semantics*, pages 17–26. Association for Computational Linguistics.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. *Journal of Artificial Intelligence Research*, pages 457–479.

Charles J. Fillmore. 1976. Frame Semantics And the Nature of Language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Luciano Floridi. 2009. Philosophical Conceptions of Information. In *Formal Theories of Information*, pages 13–53. Springer.

Satoru Fujishige. 2005. *Submodular functions and optimization*. Annals of discrete mathematics. Elsevier, Amsterdam, Boston, Paris.

Kata Gábor, Haifa Zargayouna, Isabelle Tellier, Davide Buscaldi, and Thierry Charnois. 2017. Exploring Vector Spaces for Semantic Relations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1814–1823, Copenhagen, Denmark. Association for Computational Linguistics.

Dan Gillick and Benoit Favre. 2009. A Scalable Global Model for Summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18, Boulder, Colorado. Association for Computational Linguistics.

Ben Hachey, Gabriel Murray, and David Reitter. 2006. Dimensionality Reduction Aids Term Co-Occurrence Based Multi-Document Summarization. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, pages 1–7. Association for Computational Linguistics.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring Content Models for Multi-document Summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370.

Sanda Harabagiu and Finley Lacatusu. 2005. Topic Themes for Multi-document Summarization. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 202–209.

Zellig Harris. 1954. Distributional structure. *Word*, 10:146–162.

Zhanying He, Chun Chen, Jiajun Bu, Can Wang, Lijun Zhang, Deng Cai, and Xiaofei He. 2012. Document Summarization Based on Data Reconstruction. In *Proceeding of the Twenty-Sixth Conference on Artificial Intelligence*.

Kai Hong, John Conroy, benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1608–1616, Reykjavik, Iceland.

Edwin T. Jaynes. 1957. Information Theory and Statistical Mechanics. *Physical Review*, 106:620–630.

Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. Content Selection in Deep Learning Models of Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828. Association for Computational Linguistics.

Solomon Kullback and Richard A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Victor Lavrenko. 2008. *A generative theory of relevance*, volume 26. Springer Science & Business Media.

Jure Leskovec, Natasa Milic-Frayling, and Marko Grobelnik. 2005. Impact of Linguistic Analysis on the Semantic Graph Coverage and Learning of Document Extracts. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1069–1074.

Piji Li, Lidong Bing, Wai Lam, Hang Li, and Yi Liao. 2015. Reader-Aware Multi-document Summarization via Sparse Coding. In *Proceedings of the 24th International Conference on Artificial Intelligence* , pages 1270–1276.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An Information-Theoretic Approach to Automatic Evaluation of Summaries. In *Proceedings of the Human Language Technology Conference at NAACL*, pages 463–470, New York City, USA.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 71–78.

Hui Lin and Jeff A. Bilmes. 2011. A Class of Submodular Functions for Document Summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 510–520, Portland, Oregon.

He Liu, Hongliang Yu, and Zhi-Hong Deng. 2015. Multi-document Summarization Based on Two-level Sparse Representation Model. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 196–202.

Annie Louis. 2014. A Bayesian Method to Incorporate Background Knowledge during Automatic Text Summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 333–338, Baltimore, Maryland.

Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics*, 39(2):267–300.

Hans Peter Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2:159–165.

Shulei Ma, Zhi-Hong Deng, and Yunlun Yang. 2016. An Unsupervised Multi-Document Summarization Framework Based on Neural Document Model. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1514–1523. The COLING 2016 Organizing Committee.

Inderjeet Mani. 1999. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA, USA.

Inderjeet Mani and Eric Bloedorn. 1997. Multi-document Summarization by Graph Search and Matching. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, pages 622–628, Providence, Rhode Island. AAAI Press.

Ryan McDonald. 2007. A Study of Global Inference Algorithms in Multi-document Summarization. In *Proceedings of the 29th European Conference on Information Retrieval Research*, pages 557–564.

Kathleen R. McKeown, Judith L. Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards Multidocument Summarization by Reformulation: Progress and Prospects. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, pages 453–460.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, Lake Tahoe, Nevada, USA.

Richard Montague. 1970. English as a formal language. In Bruno Visentini, editor, *Linguaggi nella societa e nella tecnica*, pages 188–221. Edizioni di Communita.

Ani Nenkova and Kathleen McKeown. 2012. A Survey of Text Summarization Techniques. *Mining Text Data*, pages 43–76.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2).

Ani Nenkova, Lucy Vanderwende, and Kathleen McKeown. 2006. A Compositional Context Sensitive Multi-document Summarizer: Exploring the Factors That Influence Summarization. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 573–580.

Maxime Peyrard and Judith Eckle-Kohler. 2016. A General Optimization Framework for Multi-Document Summarization Using Genetic Algorithms and Swarm Intelligence. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 247 – 257.

Maxime Peyrard and Judith Eckle-Kohler. 2017a. A principled framework for evaluating summarizers: Comparing models of summary quality against human judgments. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, volume Volume 2: Short Papers, pages 26–31. Association for Computational Linguistics.

Maxime Peyrard and Judith Eckle-Kohler. 2017b. Supervised learning of automatic pyramid for optimization-based multi-document summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, volume Volume 1: Long Papers, pages 1084–1094. Association for Computational Linguistics.

Maxime Peyrard and Iryna Gurevych. 2018. Objective function learning to match human judgements for optimization-based summarization. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 654–660. Association for Computational Linguistics.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies. In *Proceedings of the NAACL-ANLP Workshop on Automatic Summarization*, volume 4, pages 21–30, Seattle, Washington.

Claude E. Shannon. 1948. A Mathematical Theory of Communication. *Bell Systems Technical Journal*, 27:623–656.

Ruben Sipos, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Large-margin Learning of Submodular Summarization Models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 224–233, Avignon, France. Association for Computational Linguistics.

Karen Sparck Jones. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of documentation*, 28(1):11–21.

Victor Yakovlevich Tsvetkov. 2014. The KE Shannon and L. Floridi's Amount of Information. *Life Science Journal*, 11(11):667–671.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word Representations: A Simple and General Method for Semi-supervised Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394.

Peter D Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of artificial intelligence research*, 37:141–188.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond SumBasic: Task-focused Summarization with Sentence Simplification and Lexical Expansion. *Information Processing & Management*, 43(6):1606–1618.

Xiaojun Wan and Jianwu Yang. 2006. Improved Affinity Graph Based Multi-Document Summarization. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 181–184. Association for Computational Linguistics.

Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. 2009. Multi-document Summarization Using Sentence-based Topic Models. In *Proceedings of the ACL-IJCNLP 2009*, pages 297–300. Association for Computational Linguistics.

Warren Weaver. 1953. Recent Contributions to the Mathematical Theory of Communication. *ETC: A Review of General Semantics*, pages 261–281.

Deirdre Wilson and Dan Sperber. 2008. *Relevance Theory*, chapter 27. John Wiley and Sons, Ltd.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Recent Advances in Document Summarization. *Knowledge and Information Systems*, 53(2):297–336.

Dani Yogatama, Fei Liu, and Noah A. Smith. 2015. Extractive Summarization by Maximizing Semantic Volume. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1961–1966, Lisbon, Portugal.

Yang Zhang, Yunqing Xia, Yi Liu, and Wenmin Wang. 2015. Clustering Sentences with Density Peaks for Multi-document Summarization. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1262–1267, Denver, Colorado. Association for Computational Linguistics.

Yixin Zhong. 2017. A Theory of Semantic Information. In *Proceedings of the IS4SI 2017 Summit Digitalisation for a Sustainable Society*, 129.

Markus Zopf, Eneldo Loza Mencía, and Johannes Fürnkranz. 2016. Beyond Centrality and Structural Features: Learning Information Importance for Text Summarization. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL 2016)*, pages 84–94.

A Details about Baseline Scoring Functions

In the paper, we compare the summary scoring function θ_I against the summary scoring functions derived from several summarizers following the methodology from Peyrard and Eckle-Kohler (2017a). Here, we give explicit formulation of the baseline scoring functions.

Edmundson: (Edmundson, 1969)
Edmundson (1969) presented a heuristic which scores sentences according to 4 different features:

- **Cue-phrases**: It is based on the hypothesis that the probable relevance of a sentence is affected by the presence of certain cue words such as 'significant' or 'important'. Bonus words have positive weights, stigma words have negative weights and all the others have no weight. The final score of the sentence is the sum of the weights of its words.
- **Key**: High-frequency content words are believed to be positively correlated with relevance (Luhn, 1958). Each word receives a weight based on its frequency in the document if it is not a stopword. The score of the sentence is also the sum of the weights of its words.
- **Title**: It measures the overlap between the sentence and the title.
- **Location**: It relies on the assumption that sentences appearing early or late in the source documents are more relevant.

By combining these scores with a linear combination, we can recognize the objective function:

$$\theta_{Edm.}(S) = \sum_{s \in S} \alpha_1 \cdot C(s) + \alpha_2 \cdot K(s) \quad (14)$$

$$+ \alpha_3 \cdot T(s) + \alpha_4 \cdot L(s) \quad (15)$$

The sum runs over sentences and C, K, T and L output the sentence scores for each method (Cue, Key, Title and Location).

ICSI: (Gillick and Favre, 2009)
A global linear optimization that extracts a summary by solving a maximum coverage problem of the most frequent bigrams in the source documents. ICSI has been among the best systems in a classical ROUGE evaluation (Hong et al., 2014). Here, the identification of the scoring function is trivial because it was originally formulated as an optimization task. If c_i is the i-th bigram selected in the summary and w_i is its weight computed from D, then:

$$\theta_{ICSI}(S) = \sum_{c_i \in S} c_i \cdot w_i \quad (16)$$

LexRank: (Erkan and Radev, 2004)
This is a well-known graph-based approach. A similarity graph $G(V, E)$ is constructed where V is the set of sentences and an edge e_{ij} is drawn between sentences v_i and v_j if and only if the cosine similarity between them is above a given threshold. Sentences are scored according to their PageRank score in G. Thus, $\theta_{LexRank}$ is given by:

$$\theta_{LexRank}(S) = \sum_{s \in S} PR_G(s) \quad (17)$$

Here, PR is the PageRank score of sentence s.

KL-Greedy: (Haghighi and Vanderwende, 2009)
In this approach, the summary should minimize the Kullback-Leibler (KL) divergence between the word distribution of the summary S and the word distribution of the documents D (i.e., $\theta_{KL} = -KL$):

$$\theta_{KL}(S) = -KL(S||D) \quad (18)$$

$$= -\sum_{g \in S} \mathbb{P}_S(g) \log \frac{\mathbb{P}_S(g)}{\mathbb{P}_D(g)} \quad (19)$$

$\mathbb{P}_X(w)$ represents the frequency of the word (or n-gram) w in the text X. The minus sign indicates that KL should be lower for better summaries. Indeed, we expect a good system summary to exhibit a similar probability distribution of n-grams as the sources.

Alternatively, the Jensen-Shannon (JS) divergence can be used instead of KL. Let M be the average word frequency distribution of the candidate summary S and the source documents D distribution:

$$\forall g \in S,\ \mathbb{P}_M(g) = \frac{1}{2}(\mathbb{P}_S(g) + \mathbb{P}_D(g)) \quad (20)$$

Then, the formula for JS is given by:

$$\theta_{JS}(S) = -JS(S||D) \quad (21)$$

$$= \frac{1}{2}\left(KL(S||M) + KL(D||M)\right) \quad (22)$$

Within our framework, the KL divergence acts as the unification of *Relevance* and *Redundancy* when semantic units are bigrams.

B Proof of Theorem 1

Let Ω be the set of semantic units. The notation ω_i represents one unit. Let $\mathbb{P}_T$, and $\mathbb{P}_K$ be the text representations of the source documents and background knowledge as probability distributions over semantic units.

We note $t_i = \mathbb{P}_T(\omega_i)$, the probability of the unit ω_i in the source T. Similarly, we note $k_i = \mathbb{P}_K(\omega_i)$. We seek a function f unifying T and K such that: $f(\omega_i) = f(t_i, k_i)$.

We remind the simple requirements that f should satisfy:

- Informativeness: $\forall i \neq j$, if $t_i = t_j$ and $k_i > k_j$ then $f(t_i, k_i) < f(t_j, k_j)$
- Relevance: $\forall i \neq j$, if $t_i > t_j$ and $k_i = k_j$ then $f(t_i, k_i) > f(t_j, k_j)$
- Additivity: $I(f(t_i, k_i)) \equiv \alpha I(t_i) + \beta I(k_i)$ (I is the information measure from Shannon's theory (Shannon, 1948))
- Normalization: $\sum_i f(t_i, k_i) = 1$

Theorem 1 states that the functions satisfying the previous requirements are:

$$\begin{aligned} \mathbb{P}_{\frac{T}{K}}(\omega_i) &= \frac{1}{C} \cdot \frac{t_i^\alpha}{k_i^\beta} \\ C &= \sum_i \frac{t_i^\alpha}{k_i^\beta} \text{ , } \alpha, \beta \in \mathbb{R}^+ \end{aligned} \tag{23}$$

with C the normalizing constant.

Proof. The information function defined by Shannon (1948) is the logarithm: $I = \log$. Then, the *Additivity* criterion can be written:

$$\log(f(t_i, k_i)) = \alpha \log(t_i) + \beta \log(k_i) + A \tag{24}$$

with A a constant independent of t_i and k_i

Since $\log$ is monotonous and increasing, the *Informativeness* and *Additivity* criteria can be combined:

$\forall i \neq j$, if $t_i = t_j$ and $k_i > k_j$ then:

$$\begin{gathered} \log f(t_i, k_i) < \log f(t_j, k_j) \\ \alpha \log(t_i) + \beta \log(k_i) < \alpha \log(t_j) + \beta \log(k_j) \\ \beta \log(k_i) < \beta \log(k_j) \end{gathered}$$

But $k_i > k_j$, therefore:

$$\beta < 0$$

For clarity, we can now use $-\beta$ with $\beta \in \mathbb{R}^+$.

Similarly, we can combine the *Relevance* and *Additivity* criteria: $\forall i \neq j$, if $t_i > t_j$ and $k_i = k_j$ then:

$$\begin{gathered} \log f(t_i, k_i) > \log f(t_j, k_j) \\ \alpha \log(t_i) + \beta \log(k_i) > \alpha \log(t_j) + \beta \log(k_j) \\ \alpha \log(t_i) > \alpha \log(t_j) \end{gathered}$$

But $t_i > t_j$, therefore:

$$\alpha > 0$$

Then, we have the following form from the *Additivity* criterion:

$$\begin{aligned} \log f(t_i, k_i) &= \alpha \log(t_i) - \beta \log(k_i) + A \\ f(t_i, k_i) &= e^A e^{[\alpha \log(t_i) - \beta \log(k_i)]} \\ f(t_i, k_i) &= e^A \frac{t_i^\alpha}{k_i^\beta} x \end{aligned}$$

Finally, the *Normalization* constraint specifies the constant e^A:

$$\begin{gathered} C = \frac{1}{e^A} \\ \text{and } C = \sum_i \frac{t_i^\alpha}{k_i^\beta} \\ \text{then: } A = -\log(\sum_i \frac{t_i^\alpha}{k_i^\beta}) \end{gathered}$$

□

C Example

As an example, for one selected topic of TAC-2008 update track, we computed the $\mathbb{P}_{\frac{D}{K}}$ and compare it to the distribution of the 4 reference summaries.

We report the two distributions together in figure 2. For visibility, only the top 50 words according to $\mathbb{P}_{\frac{D}{K}}$ are considered. However, we observe

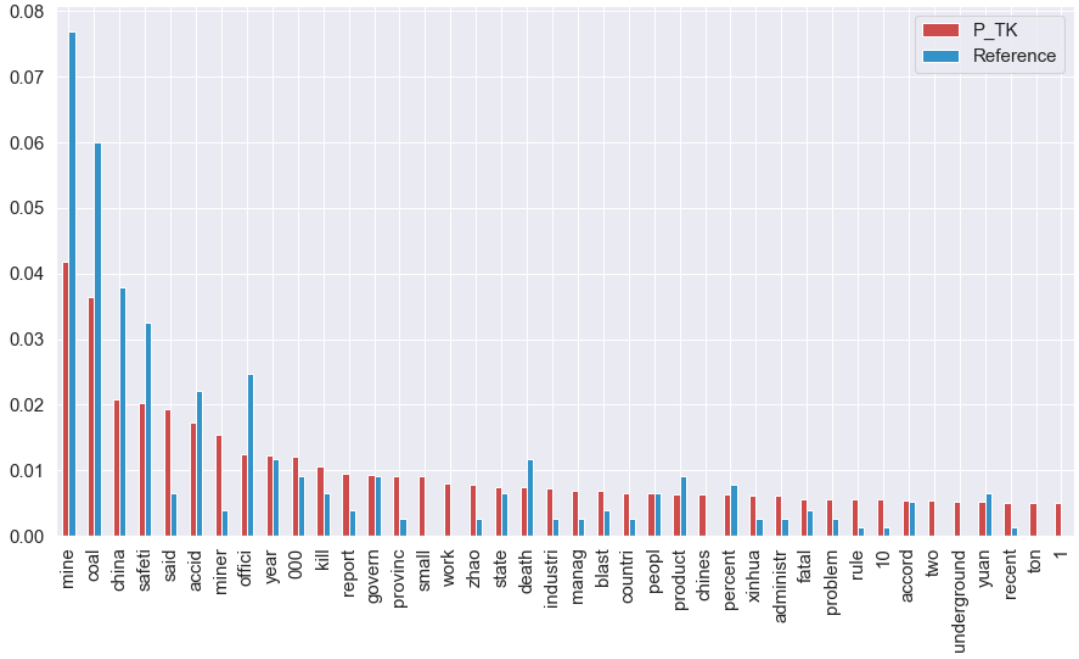


Figure 2: Example of $\mathbb{P}_{\frac{D}{K}}$ in comparison to the word distribution of reference summaries for one topic of TAC-2008 (D0803).

a good match between the distribution of the reference summaries and the *ideal* distribution as defined by $\mathbb{P}_{\frac{D}{K}}$.

Furthermore, the most desired words according to $\mathbb{P}_{\frac{D}{K}}$ make sense. This can be seen by looking at one of the human-written reference summary of this topic:

> **Reference summary for topic D0803**
> *China sacrificed coal mine safety in its massive demand for energy. Gas explosions, flooding, fires, and cave-ins cause most accidents. The mining industry is riddled with corruption from mining officials to owners. Officials are often illegally invested in mines and ignore safety procedures for production. South Africa recently provided China with information on mining safety and technology during a conference. China is beginning enforcement of safety regulations. Over 12,000 mines have been ordered to suspend operations and 4,000 others ordered closed. This year 4,228 miners were killed in 2,337 coal mine accidents. China's mines are the most dangerous worldwide.*