

On using a Quantum Physics formalism for Multi-document Summarisation

B. Piwowarski
University of Glasgow, UK
benjamin@bpiwowar.net

M. R. Amini
NRC Institute for Information Technology, Québec, Canada
massih-reza.amini@lip6.fr

M. Lalmas
Yahoo! Research Barcelona
mounia@acm.org

8th January 2012

Abstract

Multi-document summarisation (MDS) aims, for each given query, to extract compressed and relevant information with respect to the different query-related themes present in a set of documents. Many approaches operate in two steps. Themes are first identified from the set, and then a summary is formed by extracting salient sentences within the different documents of each of the identified themes. Among these approaches, Latent Semantic Analysis (LSA) based ones rely on spectral decomposition techniques to identify the themes. In this paper, we propose a major extension of these techniques that relies on the Quantum Information Access (QIA) framework. The latter is a framework developed for modelling information access based on the probabilistic formalism of quantum physics. The QIA framework allows to not only point out the limitations of the current LSA-based approaches, but motivates a new principled criterium to tackle multi-document summarisation that addresses these limitations. As a by-product, it also provides a way to enhance the LSA-based approaches. Extensive experiments on the DUC 2005, 2006 and 2007 datasets show that the proposed approach consistently improves over both the LSA-based approaches and the systems that competed in the yearly DUC competitions. This demonstrates the potential impact of quantum-inspired approaches to Information Access in general, and of the QIA framework in particular.

1 Introduction

Multi-document summarisation (MDS) systems aim to extract information relevant to an implicit or explicit query from a set of documents. These systems are commonly used in most web-oriented summarisation applications. MDS systems can be used with conventional search engines to, for instance, provide informative snippets to help users navigate through different parts of the result page (Amitay, 2001; Turpin, Tsegay, Hawking & Williams, 2007). They can also offer short summaries of documents initially clustered by a for instance news aggregator to assist users in better understanding the different views contained in the news (McKeown, Passonneau, Elson, Nenkova & Hirschberg, 2005; Sampath & Martinovic, 2002). Another application is a question & answering system which, for each asked question, supplies information about the answer in the form of a short extractive summary (Hirao, Sasaki & Isozaki, 2001).

MDS is a more complex task than single document summarisation as it aims to select sentences relevant to different query-related themes, inside a set of documents, rather than to only shorten a single source text (Lin & Hovy, 2002; Mani & Bloedorn, 1999). A major issue for MDS is to automatically detect these themes and then extract the most relevant sentences with respect to these themes to form the summary. Summaries can also be biased by the query used for searching documents. Most approaches to this task suppose that each relevant sentence to the summary must fall within one and only one of the identified themes. This assumption is too restrictive as it ignores the many candidate sentences associated with several identified themes but not specifically associated with one of them in particular.

Several Latent Semantic Analysis (LSA) based methods have been proposed for single document summarisation (Gong & Lin, 2001; Murray, Renals & Carletta, 2005; Steinberger & Ježek, 2004; Ozsoy, Cicekli & Alpaslan, 2010). In them, spectral decomposition over the vectors representing the sentences is used to detect the different themes inside a collection of documents, before selecting the sentences that are important for one or more themes, with a criterium depending on the specific LSA-based approach used. However, the different LSA-based approaches proposed so far do not compete with state-of-the-art MDS systems. One reason is that these techniques were firstly developed for single document summarisation and hence have not been optimised for MDS. The other and more fundamental reason is theoretical, and since this can be seen only within the Quantum Information Access framework (QIA), we first introduce the latter.

The QIA framework, which was originally developed in (Piwowarski, Frommholz, Lalmas & Rijsbergen, 2010) to model information retrieval, both relies on a *quantum* probabilistic theory, the quantum physics mathematical formalism, and defines a methodology to represent information objects such as textual documents. In QIA, as for LSA, extracting the salient topics of one or more documents starts by defining a set of vectors associated with sentences. The QIA framework uses this set of vectors to create a quantum probability density, i.e. a quantum distribution over vectors in a topical space.

This methodology allows first to offer a re-interpretation of the two different criteria that have been proposed in LSA and to show why they are flawed if we reformulate them within the quantum probability formalism. In addition, we propose a new criterium to select sentences for the summary that takes into account all the sentences previously selected. This criterium relies on the use of quantum events that are defined as subspaces in the topical space. Intuitively, a good summary should cover a subspace of the topical space associated with high (quantum) probability density.

To validate our QIA-based formulation of MDS we perform extensive experiments on three large date sets used in the DUC competitions, 2005 to 2007. We vary a set of parameters for both LSA and QIA based approaches (prior sentence density, weighting scheme, and rank selection in the spectral decomposition), and show that our approach consistently improves over both LSA-based summarisation techniques, and the best performing approaches in each of these competitions.

In this paper, we also report two by-products of our approach. First, we show that we can associate with each sentence a probabilistic prior, thus generalising over the proposed LSA approaches, and which, as shown in the experiments, improves over the performance of LSA approaches. Second, the QIA framework relies on a general hypothesis that, if two themes (assumed to be vectors in a topical space) are present with a non-null probability in a set of documents, then any two linear combinations of those vectors is also a theme of that collection¹. This hypothesis could not be tested easily in ad-hoc IR, but, by slightly modifying the QIA sentence selection criterium, we experimentally show in this paper that the hypothesis is not invalidated, yielding an important result for the application of the QIA framework to model information access applications.

In the remainder of the paper, we provide, in section 2, the background and motivation of our work. In section 3 we present the QIA framework and its connection with LSA-based approaches, and in section 4, we present our QIA-based approach for MDS. In section 5, we describe our experimental setup, our experimental results and their analysis. Finally, in section 6 we discuss the outcomes of this study and give pointers to further research.

¹This is discussed further in Section 4.2

2 Background and Motivation

We first present related work on MDS, before discussing and motivating the use of the QIA framework, upon which our quantum-inspired summarisation approach is based.

2.1 Multi-Document Summarisation

Research in text summarisation showed that human-quality text summarisation is very complex since it encompasses information fusion (Barzilay, McKeown & Elhadad, 1999), sentence compression (Knight & Marcu, 2002), and language generation (McKeown, Klavans, Hatzivassiloglou, Barzilay & Eskin, 1999; Sparck Jones, 1993). Simpler approaches have then be explored, consisting in extracting representative text spans, that is generating *extract summaries* instead of *abstract summaries*. Extraction approaches include statistical techniques and/or those based on surface domain-independent linguistic analysis. Within this context, query-biased MDS can be defined as the selection of a subset of sentences that is representative of topics relevant to a query or question, and present in a given collection of documents (Radev, Jing, Styś & Tam, 2004). This is typically done by ranking document sentences and selecting those with higher score and minimum overlap for each of these topics. Usually, sentences are used as text span units but paragraphs have also been considered (Mittra, Singhal & Buckley, 1997). Using the latter can be more appealing since they contain more contextual information and provide a coherent sequence of sentences. The quality of an *extract summary* might not be as good as an *abstract summary*, but it is considered sufficient enough for a reader to understand the main ideas or answers to a question. Post-processing can also be applied to produce a more coherent summary.

MDS techniques can be broadly categorised into three groups, feature-based (Harabagiu & Lacatusu, 2005; Radev et al., 2004; Amini & Usunier, 2011), graph-based (Erkan & Radev, 2004; Mihalcea, 2005; Wang, Li, Zhu & Ding, 2008) and lexical chain based (Barzilay & Elhadad, 1997; Chen, Wang & Liu, 2005; Li & Sun, 2008) methods. The former first identifies themes and then assigns scores to sentences in each of these themes based on sentence-level and inter-sentence features, e.g. sentence similarity, position, cluster centroids, etc. Graph-based techniques begin by characterising a set of documents as a weighted text graph and then recursively compute sentence significance, globally, from the entire text graph rather than using single sentences as in feature-based methods. The underlying hypothesis of both methods is that summary sentences are those belonging to an identified theme or to a sentence cluster found in the graph. Therefore, sentences relevant to more than one theme or those midway between two clusters in the graph are never extracted and hence are never part of the summary. Finally, lexical chain approaches first construct different sequences of semantically related words, chains relevant to the topic at hand are identified and eventually sentences matching these identified chains are extracted from the collection of documents.

Our proposed QIA-based approach to MDS belongs to the first group (feature-based) and bears similarity with LSA-based approaches, a group of successful approaches first proposed for single document summarisation. They aim at extracting salient sentences of a given document within a reduced term space² and are based on the singular value decomposition (SVD) of a term-sentence matrix. There are two groups of LSA-based approaches. The first (Gong & Lin, 2001; Murray et al., 2005) assumes that each topic found by SVD should be present in the final summary and select sentences having the highest entry along each of the extracted topics. Steinberger and Ježek (Steinberger & Ježek, 2004) found that sentences belonging to several “latent” topics may be good candidates for extraction but are never selected by LSA-based approaches to form the summary. To overcome this, they computed a score for each sentence that depends on the most salient extracted latent topics. As we discuss in section 3, our approach re-interpret LSA-based methods under the QIA framework, which naturally paves the way for selecting those sentences falling into one theme or more.

²Sentences are represented in a term space, and singular value decomposition (SVD) is used to find the main latent topics, i.e. the “cluster” representatives, in the original term space.

2.2 Quantum Information Access

We now turn to the QIA framework. Besides van Rijsbergen’s seminal work (Rijsbergen, 2004) who advocated for the usefulness of the quantum theory formalism in IR, studies on using quantum physics for information access have emerged to, for instance, express document ranking with the aim to capture diversity (Zuccon & Azzopardi, 2010), or to represent documents in a space different from the standard term space (Huertas-Rosero, Azzopardi & Rijsbergen, 2009). Our work is based on the Quantum Information Retrieval framework developed by Piwowarski et al (Piwowarski et al., 2010). This line of work was conducted within the remit of ad-hoc information retrieval. However, as the framework is being extended to other tasks, such as summarisation in this paper, we use the more general name of “Quantum Information Access” (QIA) to refer to this framework.

The basic assumption of QIA is that there exists a Hilbert space³ \mathcal{H} of *information needs*, called *information need space*. Taking inspiration from (Rijsbergen, 2004), QIA provided both theoretical and experimental insights on the relationships between quantum physics and information access. In this paper, we restrict ourselves to a simple information need space, namely a topical space, where each vector corresponds to a distinct topical aspect, and each dimension corresponds to a term (or a bi-gram). Such vectors are called *atomic topics*. We think such a representation is enough for the summarisation task since this latter is mainly about the detection of topics and not of other information need related spaces (such as emotion or style).

The QIA framework relies on a multi-dimensional representation of text fragments (any set of sentences), both to represent the distribution over atomic topics present in a fragment and to represent the topics covered by this fragment, by means of a subspace. Using a multi-dimensional representation of documents has been shown important in information retrieval (IR), to deal with multi-topic documents (Zuccon, Azzopardi & Rijsbergen, 2009), to build up semantic spaces (Widdows, 2004) and for contextual IR (Melucci, 2008). Among those, the work of Melucci (Melucci, 2008) is the closest to ours since it uses spectral decomposition to uncover subspaces (relevant context). Differently, in this paper, we use subspaces to represent the topics covered by an extracted summary.

Finally, as advocated in (Piwowarski et al., 2010), to our knowledge QIA is the only framework that provides a uniform and principled formalism dealing with representations of documents and information needs that span multiple dimensions. Previous works using multidimensional representation did so for either queries or documents, but not both. In this paper, we again make use of this use of multidimensional objects to represent both the “information need” (as above discussed, the topics to be discussed in the summary) and the extracted summary.

In this paper, we first analyse the LSA-based methods by showing that they can be interpreted within the QIA framework. Indeed, the scores computed to rank sentences can be shown to be (quantum) probabilities which purpose is to indicate the “goodness” of the sentence for being included in the summary. This QIA-theoretic interpretation has the advantage of clearly showing why the hypothesis of linking summary sentences exclusively to just one theme (latent topic using the LSA terminology) is flawed. We further show that under the QIA framework a more natural criterium of selecting sentences can be defined for MDS, where the latter translates into a difference in performance on the DUC test collections.

3 Quantum summarisation

In the following, we first present the QIA framework and link it with a measure on the topicality of text fragments, providing a quantitative view on the salient themes of such a set of text fragments (Section 3.1). In Section 3.2, we show the link between QIA and spectral decomposition, and then re-interpret existing LSA-based approaches within this framework in Section 3.3. Our proposed model will be presented in Section 4.

³Roughly, a vector space on the complex field with a geometric structure defined by an inner product.

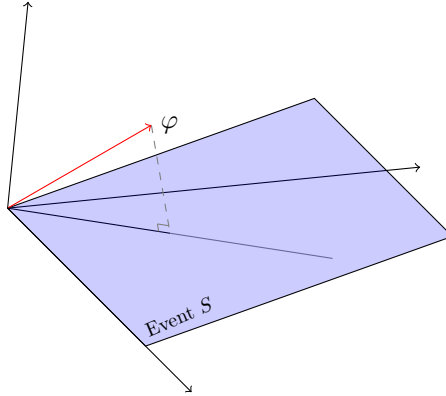


Figure 1: Quantum probabilities - The projection of φ on S

3.1 Quantum IA and summarisation

The QIA framework Quantum physics describe the behaviour of matter at atomic and subatomic scales by identifying the state of a physical system in a known state as a state vector in a Hilbert space \mathcal{H} , where a state vector is a unit vector φ in \mathcal{H} . States determine statistically the measures obtained on the system, for instance related to the position of a particle. In this case, the state vector associated with this particle determines the probability that it is at a given position.

In the QIA approach to summarisation, the concept of *system* does not refer to a physical entity, but to the topicality of a text fragment, i.e. any subpart of a set of documents. More precisely, the QIA framework (Piwowarski et al., 2010) relies on the existence of a Hilbert space \mathcal{H} of *topics*, called *topical space*, where each vector corresponds to an *atomic topic*. The latter can be compared to the notion of “factoid” (Halteren & Teufel, 2003) or “theme”, used in summarisation and question-answering to assess the amount of relevant information a summary or an answer contains. Further, a theme (vector) as extracted by LSA approaches to summarisation corresponds to an atomic topic.

An event is represented as a subspace S of the Hilbert space \mathcal{H} . In our case, a subspace can be seen as an (infinite) set of atomic topics. We can evaluate the probability that a fragment represented by the atomic topic φ “is similar” to one of the atomic topics present in the subspace⁴. If φ is strictly contained within the subspace, then the probability is 1. If φ is orthogonal to any atomic topic of the subspace, then the probability is 0. In the other cases, the closer φ is to the subspace, the closer its probability would be to 1. More formally, the probability of an event is given by the square of the length of the projection of φ onto the corresponding event subspace S , that is by computing the value $\|\hat{S}\varphi\|^2$, where \hat{S} is the projector onto the subspace S , as illustrated in Figure 1.

Note that, as evoked earlier, even when the system state is known or determined (i.e. we know which state vector φ characterise the system), the events are not certain. This is a property of the quantum physics formalism. Within the topical space, this means that even if we know the atomic topic to be φ , the probability that the text fragment deals with a topic φ' not orthogonal to φ is not null. Said otherwise, topicality is a continuum that goes from “completely not related” atomic topics (orthogonality) to “exactly the same” atomic topic (linearity).

We cannot assume that a text fragment is associated with only one atomic topic. To consider multi-topicality, we assume that a text fragment has a given probability of dealing with each atomic topic it contains, where the probability reflects the importance of each atomic topic within the text fragment. In (quantum) physics, states are exclusive, i.e. a system can be in only one state at any given time. Similarly,

⁴Here, “similarity” is to be interpreted both as the standard cosine similarity of IR (intuitive point of view) and as a quantum probability (theoretical point of view). It is the quantum view that is described in this paragraph.

we can imagine that each text fragment has an associated set of atomic topics, and each time we want to measure the topicality of the fragment, we pick one of these only. As states are mutually exclusive, following standard probability theory, we require that the probability over the atomic topics sums up to 1. Thus, given a probability distribution over the topics $p(\varphi)$, we define the probability of an event S , where S means “the text fragment is about the topics defined by S ”, as:

$$q(S) = \sum_{\varphi} p(\varphi) q(S|\varphi) = \sum_{\varphi} p(\varphi) \left\| \widehat{S}\varphi \right\|^2 \quad (1)$$

We use the symbol q to denote the quantum probability measure. Note that the above equation reduces to $\left\| \widehat{S}\bar{\varphi} \right\|^2$ if $p(\cdot)$ is null for all φ except the vector $\bar{\varphi}$, i.e. when there is no uncertainty about the topical state.

This probability is also “*quantum*”, i.e. it does not obey standard probability laws. This can be seen easily by showing that the sum of the probabilities of three mutually exclusive events is superior to 1. To illustrate this, consider the three events associated with the one-dimensional subspaces S_1 and S_2 , respectively, associated with the vectors φ_1 and φ_2 in Figure 2. If the probability distribution is defined by $p(\varphi_1) = 1$, then $q(S_1) = 1$ and $q(S_2) = (\varphi_1 \cdot \varphi_2)^2 > 0$. The sum of both is indeed strictly greater than 1.

Representing the topicality of text fragments We describe now how the QIA framework is used to represent text fragments. The representation is based on two assumptions: (1) a fragment typically contains various atomic topics; and (2) each fragment can be split into (possibly overlapping and non-contiguous) different *atomic* fragments, where each atomic fragment addresses one atomic topic. This follows from research in focused retrieval, where answers to a query usually correspond to document excerpts (sentences or paragraphs) and not full documents (Piwowarski, Trotman & Lalmas, 2009).

In this paper, following the extractive summarisation literature, we assume that the atomic topics are in one-to-one relationship with sentences, i.e. that each sentence is an atomic fragment. Even though in an ad-hoc IR sliding windows over the text yielded better results (Piwowarski et al., 2010), we chose to keep sentences as atomic fragments for two reasons. First, texts used for summarisation in DUC are news articles and not web pages as in some e.g. of the TREC collections, and hence sentence extraction algorithms are performing better. Second, sentences are a natural unit in extractive summarisation and were used by all other LSA-based techniques.

From an intuitive point of view, it should be noted that using sentences or sliding windows is not fully satisfactory, and that sentences generally map to more than one atomic topics or factoids (Halteren & Teufel, 2003). In theory, it would be useful to be able to extract and represent such factoids, but in practice both of these problems are complex. In this paper, we adopt a simpler approach where sentences are atomic topics and the text they contain is used straightforwardly to represent the corresponding atomic topic.

A fragment \mathcal{F} is then identified by the sequence $\varphi_1, \dots, \varphi_f$ of f atomic topic vectors corresponding to the f sentences of the fragment. We also denote φ_s the atomic vector associated with the sentence s in the fragment \mathcal{F} .

Text fragments can be represented in two ways using QIA, as a distribution of probability over atomic topics, or as an event corresponding to the atomic topics present within the fragment. In order to do so, we first need to define a probability distribution over the f sentences of a fragment \mathcal{F} .

In the most general case, we assume that the set of atomic topics corresponding to a fragment can only be a subset of those that appear in the fragment. In practice, we give a prior $p(s|\mathcal{F})$ to the probability that the sentence s represents the fragment atomic topic. Using the Kronecker delta function $\delta_{\varphi\varphi_i}$ (which is equal to 1 if and only if φ coincides with φ_i and 0 otherwise), this gives:

$$p(\varphi|\mathcal{F}) = \sum_{s \in \mathcal{F}} p(s|\mathcal{F}) \delta_{\varphi\varphi_s} \quad (2)$$

The most straightforward way to define the prior $p(s|\mathcal{F})$ over sentences is to assume that all sentences are equally important, so the distribution over the sentences is uniform, i.e.

$$\forall s \in \mathcal{F}, p_0(s|\mathcal{F}) = \frac{1}{\text{number of sentences}} \quad (3)$$

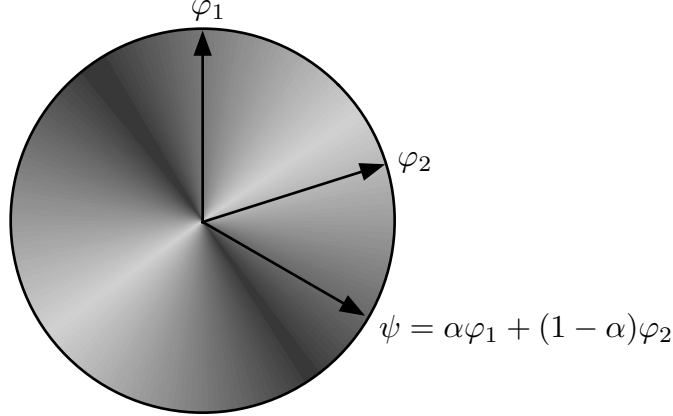


Figure 2: Illustration of a density in two dimensions – darker areas mean higher probability. In this figure, we can see that the probability density smoothly changes with respect to normalised linear combinations of vectors.

This is the approach (implicitly) taken by all the LSA approaches for MDS. We present in Section 4.1 other priors that perform better experimentally, and can also be used within LSA approaches, when interpreted within our QIA framework.

Equations 1 and 2 define a quantum probability distribution, $q(S|\mathcal{F})$:

$$q(S|\mathcal{F}) = \sum_{\varphi} p(\varphi|\mathcal{F}) \left\| \hat{S}\varphi \right\|^2$$

The above allows us to illustrate the fundamental hypothesis upon which the QIA framework relies. Let us consider the case of the simplest type of events, i.e. one-dimensional subspaces S_{φ_i} defined by a vector φ_i . If one or both events S_{φ_1} and S_{φ_2} have a non-null probability, then any event S_{ψ} associated with a linear combination of these two vectors has also a non-null probability. This is illustrated in Figure 2 and can be shown using Equation 4, given in the next section. In our experiments, we show that the QIA hypothesis is not invalidated. To prove that the hypothesis holds is in practice impossible, since we would have to prove it on a theoretical basis, and Information Access is (mostly) experiment-driven; hence we can only show through experimental evidence that the hypothesis “holds”⁵.

We discuss now the second possible representation of a fragment, as an event corresponding to the topics covered by it. We assume that the subspace corresponding to the fragment should contain each atomic topic in the fragment with a probability of 1, i.e. that $q(S_{\varphi_s}|\mathcal{F})$ equals 1 for any sentence s in the fragment \mathcal{F} . As discussed above, any linear combination of two atomic topic vectors has an associated non-null probability. Consequently, the subspace corresponding to a fragment is the span of the different vectors in \mathcal{F} . We denote $S_{\mathcal{F}}$ the subspace associated with a fragment \mathcal{F} .

This dual view of the topicality of text fragments, and more generally of information objects, is at the core of the QIA framework and is used when interpreting the LSA-based approaches in the two next sections, as well as when we define our proposed criterium for summarisation.

3.2 Spectral decomposition and QIA

To link the proposed approach to LSA-based ones (as described in the next section), we first need to relate the QIA framework with spectral decomposition. To this end, we first derive a computable version of the (quantum) probability $q(S|\mathcal{F})$, following the usual approach taken in Quantum Physics of using the trace operator (Nielsen & Chuang, 2000):

⁵This experiment-driven “proofs” of hypotheses can be found in many other works in IR, and more particularly in works working on the axiomatic of IR, e.g. (Fang, Tao & Zhai, 2011).

$$\begin{aligned}
q(S|\mathcal{F}) &= \sum_{\varphi} p(\varphi|\mathcal{F}) \left\| \widehat{S}\varphi \right\|^2 = \sum_{\varphi \in \mathcal{V}} \text{tr} \left(p(\varphi|\mathcal{F}) \varphi^\top \widehat{S} \varphi \right) \\
&= \text{tr} \left(\widehat{S} \underbrace{\sum_{\varphi} p(\varphi|\mathcal{F}) \varphi \varphi^\top}_{\rho_{\mathcal{F}}} \right) = \text{tr}(\widehat{S} \rho_{\mathcal{F}})
\end{aligned} \tag{4}$$

where $\rho_{\mathcal{F}}$ is a probability density operator, a terminology coming from the quantum formalism. It can be shown that any positive semi-definite linear operator ρ of trace 1 is a valid probability density operator (Rijsbergen, 2004). The interest of this reformulation is that we have a product of two linear operators, i.e. matrices, \widehat{S} and $\rho_{\mathcal{F}}$, which, respectively, corresponds to the event (subspace) and the density operator.

From Equations 2 and 4, the density associated with \mathcal{F} is

$$\rho_{\mathcal{F}} = \sum_{s \in \mathcal{F}} p(s|\mathcal{F}) \varphi_s \varphi_s^\top \tag{5}$$

where φ_s is the atomic vector associated with the sentence s in the fragment \mathcal{F} . As any self-adjoint linear operator of finite rank, the density $\rho_{\mathcal{F}}$ can be decomposed, using eigenvalue decomposition, into

$$\rho_{\mathcal{F}} = U \Sigma^2 U^\top$$

where U is an orthonormal matrix and Σ is a diagonal matrix of non-null eigenvalues. This defines the spectral decomposition of the density view on fragments, i.e. of the density associated with a fragment \mathcal{F} .

Note that the lowest eigenvalues are usually discarded since they correspond to meaningless dimensions, i.e. dimensions associated with noise (Deerwester, Dumais, Furnas & Landauer, 1990), which is in our case due the process of extracting atomic topics from text. The k^{th} rank approximation of A can be written

$$\rho_{\mathcal{F}}^{(k)} = U^{(k)} \Sigma^{2(k)} V^{(k)\top}$$

where $U^{(k)}$ and $V^{(k)}$ are restrictions of U and V respectively to their k first columns, and $\Sigma^{2(k)}$ corresponds to the first k columns and rows of Σ^2 .

We now turn to the second view on fragments, that of a subspace/projector. From the above decomposition, we can define the projector $\widehat{S}_{\mathcal{F}}$ associated with the subspace spanned by the atomic topic vectors of fragment \mathcal{F} . There, the columns of U form the basis of the subspace that contains any linear combination of the atomic vectors, and hence it can be shown that

$$\widehat{S}_{\mathcal{F}} = U U^\top$$

As in the case of the eigenvalue decomposition, we use only the first k columns of U to discard dimensions associated with noise:

$$\widehat{S}_{\mathcal{F}}^{(k)} = U^{(k)} U^{(k)\top}$$

We showed how the density and the projector (associated with the subspace) can be represented using eigenvalue decomposition. This provides the necessary basis for the derivations connecting QIA to the LSA-based summarisation, which we describe next. To de-clutter notations, we drop (k) in the remaining of the paper.

3.3 Connections with LSA summarisation

In this section, we link QIA as described above, with the LSA-based summarisation techniques. We focus on two techniques, that of Gong and Lin (Gong & Lin, 2001) and Steinberger and Ježek (Steinberger & Ježek, 2004), since all others are variations of them. We adapt the notations for clarity.

LSA-based techniques are based on the singular value decomposition (SVD) of the term-sentence matrix A , where each column is associated with a sentence from the set of documents \mathcal{D} to summarise and each row to a distinct term:

$$A = U\Sigma V^\top \quad (6)$$

where U and V are orthonormal matrices and Σ is a diagonal matrix with decreasing entries $\sigma_1 < \dots < \sigma_n$. Each singular value σ_i corresponds to what we call in this paper an SVD atomic topic⁶. The columns of U represent the atomic topics in the term space. The columns of V represent the atomic topics in the sentence space, i.e. the magnitude of the matrix entry V_{ij} corresponds to the importance of sentence i for atomic topic j (Gong & Lin, 2001).

Without loss of generality, we assume that each column of A has a norm equal to the inverse of the number of sentences in the set of documents \mathcal{D} . This allows us to link this SVD decomposition to the previous section and hence to the QIA framework. More precisely, by assuming that the distribution over sentences $p(s|\mathcal{D})$ is uniform, we can then write, using Equation 5,

$$\rho_{\mathcal{D}} = \sum_{s \in \mathcal{D}} p(s|\mathcal{D}) \varphi_s \varphi_s^\top = AA^\top \quad (7)$$

which also implies that $\rho_{\mathcal{D}}$ equals to $U\Sigma^2U^\top$.

Using the above, we now show how the two LSA-based techniques above mentioned can be expressed within the QIA framework. We use $X_{\bullet j}$ (respectively $X_{i\bullet}$) as a short-hand for the j^{th} column (respectively i^{th} row) of a matrix X . We denote s_i the i^{th} sentence of the set of documents, i.e. the sentence corresponding to the i^{th} column of A .

To form a summary, Gong and Lin (Gong & Lin, 2001) use the k atomic topics associated with the k highest singular values⁷, i.e. with $\sigma_1, \dots, \sigma_k$. The j^{th} atomic topic is represented in the sentence space by the j^{th} column of the matrix V (Equation 6). The i^{th} entry V_{ij} of this vector corresponds to the importance of the i^{th} sentence for the j^{th} atomic topic. Formally, for the j^{th} atomic topic, Gong and Lin (Gong & Lin, 2001) select the i_*^{th} sentence such that:

$$i_* = \operatorname{argmax}_i V_{ij}^2$$

Using the fact that $V = A^\top U \Sigma^{-1}$, we can rewrite this selection criterium as:

$$\begin{aligned} \operatorname{argmax}_i V_{ij}^2 &= \operatorname{argmax}_i (A^\top U \Sigma^{-1})_{ij}^2 = \operatorname{argmax}_i (A^\top U)_{ij}^2 \Sigma_{jj}^{-1} \\ &= \operatorname{argmax}_i (s_i^\top U_{\bullet j})^2 = \operatorname{argmax}_i \operatorname{tr}(U_{\bullet j} U_{\bullet j}^\top s_i s_i^\top) \\ &= \operatorname{argmax}_q \left(\mathcal{S}_{\mathcal{D}}^{(j)} | s_i \right) \end{aligned} \quad (8)$$

where $\mathcal{S}_{\mathcal{D}}^{(j)}$ is the one-dimensional subspace associated with the j^{th} column of U , i.e. to the j^{th} latent atomic topic. Hence, the selection process corresponds to maximising the probability associated with the j^{th} dimension of the subspace $\mathcal{S}_{\mathcal{D}}$ that represent the salient topics of the documents to summarise. This means that a sentence that is a combination of two atomic topics (j_1) and (j_2) might not be selected because it lies half way between the subspaces $\mathcal{S}_{\mathcal{D}}^{(j_1)}$ and $\mathcal{S}_{\mathcal{D}}^{(j_2)}$. However, this topic, according to the hypotheses of the QIA

⁶The standard terminology in summarisation is a latent topic, or SVD theme.

⁷If there are less than k non-null singular values, the method cycles through the singular values, beginning with the highest ones.

framework, is fully contained with the topics of the documents, and would constitute a good candidate for the summary.

This is an illustration of the problem of the hard clustering existing in Gong and Lin selection method. This problem is further exacerbated when singular values are close to each other. In the extreme case where they are equal, i.e. σ_{j_1} and σ_{j_2} , the SVD problem is degenerate, i.e. the two vectors can be any two that define the same two-dimensional subspace, making the criterium arbitrary and sensitive to numerical approximations.

Steinberger and Ježek (Steinberger & Ježek, 2004) also noticed this problem. Although they did not give a principled explanation of the underlying reason, they noted that a sentence can be highly ranked for many atomic topics but never sufficiently to be selected. The approach they proposed is to first select an appropriate rank k for approximation of the matrix A . Then, they proposed to select the i^{th} sentence that maximises the following criterium:

$$g_i = \sum_{j=1}^k V_{ij}^2 \sigma_j^2 = \text{tr} (V_{i\bullet} \Sigma^2 V_{i\bullet}^\top)$$

Since $V_{i\bullet}$ equals $s_i^\top U \Sigma^{-1}$, we have

$$g_i = \text{tr} (s_i^\top U U^\top s_i) = q(\mathcal{S}_{\mathcal{D}} | s_i) \quad (9)$$

where s_i is a pure atomic topic state, i.e. we know that the atomic topic is s_i . Hence, this criterium select sentences maximising the probability of being present in the most important (i.e. k) document topics.

This method has two shortcomings. First, it assumes that the dimension of $\mathcal{S}_{\mathcal{D}}$ is correctly chosen, because if the rank is maximal the probability defined by Equation 9 is always equal to 1 since $\mathcal{S}_{\mathcal{D}}$ is a subspace that contains all the atomic vectors present in the documents of \mathcal{D} . Second, different to (Gong & Lin, 2001), sentences close to only one SVD atomic topic can be selected repeatedly. While for important atomic topics, i.e. those with high singular values, this can be a good property, it may lead to too much homogeneity in the summary. In the worst case, a sentence that occurs more than one time in the documents to summarise can be chosen repeatedly.

In the next section, we propose an approach that cater for atomic topics that (1) are combination of the SVD atomic topics, hence overcoming Gong and Lin (Gong & Lin, 2001) problems, and that (2) extract sentences from different topics, hence overcoming the limitations of Steinberger and Ježek (Steinberger & Ježek, 2004).

4 The QIA-based approach

In this section, taking advantage of the quantum probability framework, we first describe alternatives to the uniform sentence prior discussed in Section 3.1 (Equation 3). We then go further, and describe our approach for MDS based on QIA. More precisely, we propose a measure of the ‘‘summarisation quality’’ of a set of sentences that is linked to how much of the probability mass of atomic topics in the documents to be summarised is covered. We also demonstrate from a theoretical perspective that the proposed measure, which is motivated by the quantum formalism, has none of the disadvantages listed in the previous section.

4.1 Sentence prior

In this section, we define the *importance* of each sentence from the documents to be summarised by setting the prior probability $p(s|\mathcal{D})$ of a sentence s defined by Equation 2. In our case, the importance should correspond to the likeliness that the atomic vector associated to sentence s be discussed within the summary. To define quantitatively how *important* is a sentence, we consider the four following prior distributions over the sentences of the documents to be summarised:

1. The uniform prior p_0 (Equation 11);

2. The document uniform prior p_d , which accounts for the varying number of sentences in each document (Equation 12);
3. The topic-biased prior p_t , which depends on the presence of query terms in the sentence (Equation 13);
4. The length-biased prior p_l , which accounts for the varying length of sentences (Equation 14).

We use a parameterised mixture of these distributions to form the final prior $p(s|\mathcal{D})$ on sentences:

$$p(s|\mathcal{D}) = \alpha_0 p_0(s|\mathcal{D}) + \alpha_d p_d(s|\mathcal{D}) + \alpha_t p_t(s|\mathcal{D}) + \alpha_l p_l(s|\mathcal{D}) \quad (10)$$

where α_\bullet are positive real values summing to 1. We describe each of the prior probabilities next.

4.1.1 Uniform prior

The initial prior p_0 defines the importance of a given sentence, regardless of its length, its relationship with the topic or of the number sentences in the document. It assumes that all sentences are equally important:

$$p_0(s) = \frac{1}{\sum_{d \in \mathcal{D}} \# \text{sentences}(d)} \quad (11)$$

where $\# \text{sentences}(d)$ is the total number of sentences in the document d .

4.1.2 Document uniform prior

The previous prior gives more importance to longer documents since the probability of selecting a sentence from a given document is directly proportional to the number of sentences it contains. An alternative approach is to consider that each document is as important as another, i.e. we first sample documents with a uniform probability of $1/\text{card}(\mathcal{D})$. We then assume that within a document, there is an equal chance that the important topics be defined by any of the sentences present in the document. Given these assumptions, we can write the distribution over the sentences given the set of documents \mathcal{D} :

$$p_d(s|\mathcal{F}) = \frac{1}{\text{card}(\mathcal{D})} \times \frac{1}{\# \text{sentences}(d_s)} \quad (12)$$

where d_s is the document containing the sentence s .

4.1.3 Topic-biased prior

This prior depends directly on the topic keywords. We chose to define it as a probability $p_t(s)$ that corresponds to the probability of picking the sentence s if we select by random a sentence containing an occurrence of any of the topic keywords. This gives

$$p_t(s) = \frac{\# \text{topic terms}(s)}{\# \text{topic terms}(\mathcal{D})} \quad (13)$$

where $\# \text{topic terms}(\bullet)$ is the number of topic terms present in the sentence s or the set of documents \mathcal{D} (the number includes the repetition of the topic terms).

4.1.4 Length-biased prior

So far sentences of various lengths have all the same importance, but in summarisation it is known that short or long sentences should not be part of summaries, and might hence not be good candidates for important atomic topics. We chose to follow an approach where we first suppose that the distribution of lengths follows a normal distribution $\mathcal{N}(\mu, \sigma)$, and estimate the maximum likelihood mean and variance using the set of documents of the summary. We then defined the prior p_l as

$$p_l(s) \propto \mathcal{N}(\text{length}(s); \mu, \sigma) \quad (14)$$

that is, the length prior is proportional to the density distribution over lengths of sentences. In that way, we give a higher prior to sentences that are of average length.

4.1.5 “Back-porting” to LSA approaches

The fact that the QIA framework relies on a probabilistic theory makes explicit how normalisation can be used for MDS. It also becomes possible to port these normalisation techniques back in the LSA approaches, thus providing the means for these approaches to benefit from new normalisation schemes derived from the QIA approach to summarisation.

In order to do so, let us consider the term-sentence matrix A . According to Equation 7, each column of this matrix corresponds to the representation of the atomic topic of the corresponding sentence multiplied by the prior sentence probability. Hence, we should normalise the i^{th} column of A so its norm equals to $p(s_i|\mathcal{D})$ in order to use the QIA prior in LSA-based approaches.

4.2 Selection criteria

We have now defined the distribution of atomic topics of a set of documents. The next step is to define how to select the sentences that will form the summary. To this end, we propose to optimise the probability that an atomic topic of a document is contained into the atomic topics of the summary. With this view, the summarisation task, as investigated in this paper, can then be stated as the following optimisation problem:

<p>Find the set of sentences $\{s_1, \dots, s_n\}$ such that</p> $S^* = \operatorname{argmax}_{s_1, \dots, s_n} q(\mathcal{S}_{s_1, \dots, s_n} \mathcal{D}) \quad (15)$ <p>where $\mathcal{S}_{s_1, \dots, s_n}$ is the subspace spanned by the atomic vectors associated with sentences s_1, \dots, s_n, and the probability $q(\mathcal{S} \mathcal{D})$ is defined by Equations 4 and 10.</p>

This optimisation overcomes the limitations of Gong and Lin (Gong & Lin, 2001), since a sentence can be selected even if it does not match an SVD atomic topic. It also addresses the limitations of Steinberger and Ježek (Steinberger & Ježek, 2004), since it would discard similar sentences that do not increase the dimensionality of the subspace $\mathcal{S}_{s_1, \dots, s_n}$.

As optimising over a set of sentences is computationally intractable, we employed two greedy approaches, where sentences are selected one by one.

4.2.1 Greedy approach 1 (QIA-1)

As a first approach, we try at each step to select the sentence s_n^* that maximises the criterion given by Equation 15 if added to an already constructed set of sentences s_1^*, \dots, s_{n-1}^* . That is, s_n^* is given by

$$s_n^* = \operatorname{argmax}_s q(\mathcal{S}_{s_1^*, \dots, s_{n-1}^*, s} | \mathcal{D})$$

In practice, we use an equivalent but computationally more efficient criterion, based on the projection of the vector φ_s , which is the atomic topic corresponding to the sentence s , onto the subspace $\widehat{\mathcal{S}}_{n-1}^\perp$ defined as the orthogonal of $\mathcal{S}_{s_1^*, \dots, s_{n-1}^*}$:

$$\begin{aligned}
s_n^* &= \operatorname{argmax}_s q \left(\frac{\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s}{\|\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s\|} \left(\frac{\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s}{\|\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s\|} \right)^\top \middle| \mathcal{D} \right) \\
&= \frac{\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s}{\|\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s\|} \rho_{\mathcal{D}} \left(\frac{\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s}{\|\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s\|} \right)^\top
\end{aligned} \tag{16}$$

Intuitively, we measure the probability that the new dimension of the subspace brought by the vector φ_s , that is the vector

$$\frac{\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s}{\|\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s\|}$$

matches the salient atomic topics of the set of documents \mathcal{D} .

4.2.2 Greedy approach 2 (QIA-2)

The second approach was not designed to improve over the previous one, but to allow us to test whether one of the hypothesis of the QIA framework holds, namely the fact that if two atomic topic vectors are contained within a fragment, then the fragment is also about any atomic topic made of the linear combination of these.

To investigate this, we notice that in Equation 16, the normalisation factor $\|\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s\|^{-1}$ ensures that the projected vector $\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s$ has a unit norm. By discarding this normalisation factor, we modify the criterium so that it “discounts” vectors that are not orthogonal to the subspace \mathcal{S}_{n-1} :

$$s_n^* = \operatorname{argmax}_s q \left(\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s \left(\widehat{\mathcal{S}}_{n-1}^\perp \varphi_s \right)^\top \middle| \mathcal{D} \right) \tag{17}$$

According to the QIA hypothesis, adding φ_s should add up a new dimension if φ_s does not belong to the subspace \mathcal{S}_{n-1} , whether φ_s is completely or only “partially” orthogonal to the subspace \mathcal{S}_{n-1} .

This is illustrated by Figure 3 where the plane is \mathcal{S}_{n-1} . The QIA hypothesis states that there is no difference between choosing φ_1 or φ_2 , which is enforced by QIA-1. That is important to choose a vector that expands the subspace in the right dimension. Unfortunately, this vector cannot be represented in three dimensions, but the reader can imagine that both vectors have a fourth component which differs, while the fourth component of the plane is set to 0. While both are orthogonal to the plane, the difference in this fourth dimension is of importance, but not the orthogonality of the vectors to the plane. In the case of QIA-2, we also take into account “how” orthogonal to the plane the vectors are, hence defining a heuristic criterium that in practice “ignores” the QIA fundamental hypothesis. If QIA-2 performs significantly better than QIA-1, then the QIA hypothesis is either false or the chosen representation of sentences is wrong.

4.3 Summary of the QIA approach

In this section, we described the QIA-based approach to MDS. The approach is defined by a general criterium (Equation 15) measuring how a subspace in the topical space, defined by the extracted sentences, covers the high probability density regions of the topical space. To define the topical density, we use the set of vectors that represent sentences from the documents to summarise, and we associate with each a given prior probability (Section 4.1).

For computational purposes, we defined, based on the general criterium, a first greedy criterium (QIA-1) that selects sentences one by one. This criterium was slightly modified (QIA-2) to investigate whether the QIA framework hypothesis is invalidated or not.

As a by-product of our approach, we also discussed how the prior over sentences can be “back-ported” to the LSA approaches, namely the rank selection and the prior over sentences (Section 4.1.5).

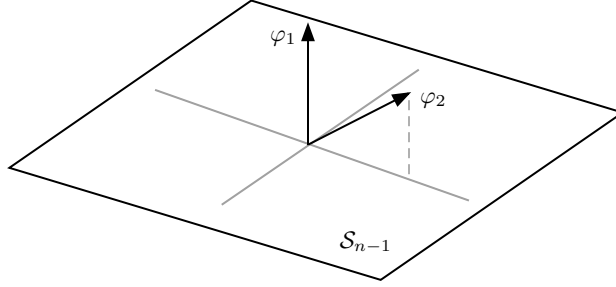


Figure 3: Illustration of the two greedy approaches for QIA-based summarisation

5 Experiments

In this section, we report the experiments conducted to validate our QIA approach. Since we introduced not only a new criterium, but also a sentence prior and (as described latter) a series of parameters, we optimise parameters for both QIA and LSA-based approaches. This allows us to see to which extent it is the criterium or the new parameters that affect performance. All experiments can be reproduced using the DUC document collections and evaluation tools, and the open-source source code of the QIA project⁸.

From now on, we use *model* to refer to one of the four LSA or two QIA based approaches. We use *system* to refer to a model with a specific set of parameters. The plan of this section is as follows. In Section 5.1, we define the collection and metrics we used for our experiments. In Section 5.2, we define our experimental set-up, which led two sets of results. Sections 5.3 to 5.5 report on the optimisation of the parameters for the different models whereas the Section 5.6 reports the final results we obtained with the optimised models.

5.1 Collection and metrics

We conducted our experiments on the DUC 2005 to 2007 data sets⁹. Documents consist of news articles collected from TREC for DUC 2005 and the AQUAINT corpus for DUC 2006 and 2007. We were interested in the main task¹⁰ of DUC 2007, i.e. providing a summary of no more than 250 words for each topic to answer the associated question. For a given question, a summary is to be formed on the basis of a subset of documents to its corresponding topic. Table 1 contains a description of the three data sets.

For each topic, we have three reference summaries produced by human assessors, which are used for evaluation. The topic questions in DUC 2005 contain in average one additional term than those in DUC 2006 and DUC 2007. In addition, the average number of terms is higher in DUC 2006 than in the two other collections. Moreover, in all three collections, the average size of sentences containing question terms (denoted by \mathbf{q} in Table 1) is 8 to 9 words higher than the average size of sentences not containing these terms.

To compare the performance of the systems, we used the ROUGE (Lin, 2004) toolkit (version 1.5.5) used by NIST for performance evaluation. This toolkit measures the quality of a produced summary by counting the relative number of unit overlaps with a set of reference summaries – in our case, those produced by three human assessors. The most employed ROUGE measure is ROUGE- n defined as:

$$\text{ROUGE} - n = \frac{\sum_{C \in \mathcal{R}} \sum_{n_{gram} \in C} \text{Count}_{\text{match}}(n_{gram})}{\sum_{C \in \mathcal{R}} \sum_{n_{gram} \in C} \text{Count}(n_{gram})}$$

⁸<http://qir.sourceforge.net>

⁹<http://www-nlpir.nist.gov/projects/duc/data.html>

¹⁰We ignored the short summary task (less than 100 words), which was abandoned in 2008 because of its difficulty for extractive summarisation methods.

	DUC 2005	DUC 2006	DUC 2007
Data source	TREC	AQUAINT	AQUAINT
Task	–	–	main
# of topics	50	50	45
# of relevant docs. per topic	25 – 50	25	25
Avg. # of keywords per topic	3.94	4.34	3.71
Avg. question size (in words)	12.42	11.26	11.35
Avg. sentence size (in words)	(q) 28.11 (-) 19.97	29.3 21.47	28.23 20.66
Summary length (in words)	250	250	250
# of participants	31	34	31

Table 1: Data sets characteristics

where \mathcal{R} is the set of reference summaries, n is the length of the n -gram, $Count_{match}(n_{gram})$ is the number of n -grams co-occurring in a produced summary and the reference summaries and $Count(n_{gram})$ is the number of n -grams in the reference summaries. In practice the overlapping units used in DUC evaluations are either unigrams or bigrams (i.e. $n \in \{1, 2\}$). ROUGE-1 score has been shown to mostly correlate with human judgments (Lin & Hovy, 2003). Other evaluation metrics implemented in ROUGE include ROUGE-L, ROUGE-W and ROUGE-SU4. ROUGE-L considers the longest common subsequence between the produced summary and the reference summaries, whereas ROUGE-W is a weighted version of the latter with usually $W = 1.2$. Finally, ROUGE-SU4 uses bi-grams with a maximum distance of four between the two words defining the bi-gram.

The ROUGE toolkit generates recall, precision and F-measure scores for all the above ROUGE metrics. In this paper, we use the average F-measure scores for ROUGE-2 and ROUGE-SU4 as it was used in DUC competitions.

5.2 Experimental setup

Documents to summarise were pre-processed by first segmenting sentences using a script¹¹ provided by NIST for DUC. All terms were converted to lowercase, digits were mapped to a single digit token, and non alpha-numeric characters were suppressed. We also used a stop-list to remove very frequent words¹².

We conducted a number of experiments aimed at evaluating how our proposed models performed in comparison to all the existing LSA-based models to summarisation we identified, and evaluating the impact of different parameters on those methods. The models that we compared with are the following (we use the name of the first author to characterise each model):

Gong Gong and Lin (Gong & Lin, 2001) model was the first LSA-based approach for text summarisation, and is described by Equation 8;

Murray Murray, Renals, and Carletta (Murray et al., 2005) model is based on a modification of the Gong and Lin model, where atomic topics are sampled according to the magnitude of their corresponding eigenvalues, i.e. the number of sentences selected with respect to one atomic topic is proportional to its corresponding eigenvalue;

Steinberger We used the criterium proposed by Steinberger and Ježek (Steinberger & Ježek, 2004) approach (Equation 9);

¹¹<http://duc.nist.gov/duc2004/software/duc2003.breakSent.tar.gz>

¹²<http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop>

	Name	Possible values
	Model	The model used among QIA (Greedy-1 and Greedy-2), Gong (Gong & Lin, 2001), Murray (Murray et al., 2005), Steinberger (Steinberger & Ježek, 2004) and Ozsoy (Ozsoy et al., 2010)
5.3	Density Rank	How the rank of the density was selected.
	Subspace Rank	How the rank of the subspace was selected (only for QIA-based approaches).
5.4	Indexed units	Uses unigrams , bigrams or both. In the case of bigrams, they can be strict or not (i.e. separated by stopped words).
	Weighting scheme	term frequency (tf), and in the case of unigram indexed units, term frequency - inverse document frequency (tf-idf) or normalised (zero mean and unit variance)
	Part-of-speech (POS) filter	Restrict to noun/verbs part-of-speech (NN,NNS,NP and NPS categories) or not .
5.5	Prior weights α_0 , α_d , α_l and α_t	Weights for the document (α_0), length (α_l), topic (α_t) and the document (α_d) priors as defined in Equation 10. By default, $\alpha_0 = 1$ and the remaining weights are set to 0.

Table 2: Summary of the different parameters used in the experiments for the different LSA and QIA based models. Sections where the different settings were experimented are shown in the leftmost column. Values in bold were those used by default and correspond to the different LSA approaches parameters in the literature.

Ozsoy We used the *Cross* method described in the work of Ozsoy, Cicekli and Alpaslan (Ozsoy et al., 2010), which is a variation of the model of Steinberger and Ježek. The authors proposed to first compute the mean value of a sentence to belong to a topic (row of matrix V^T), and then set to zero all the values below this mean value, hence defining a threshold below which a sentence is not at all considered to be discussing an atomic topic before following the approach of Steinberger and Ježek.

In all these models, the parameter to set is the rank of the decomposition, i.e. the rank of the density, which corresponds to the first optimisation we make (Section 5.3). However, we go further and experiment with a range of parameters summarised in Table 2. Due to the high number of parameters, we optimise their values for each model following three steps:

1. In Section 3.2, we discussed the problem of noise and its relationship with the selection of an appropriate rank of the selection. A rank selection method is needed when computing the density $q(\cdot|\mathcal{D})$, or the subspace \mathcal{S}_n . We experimented with the following strategies

None (Only for subspaces) No rank selection was applied;

Mean We selected the eigenvalues above the average of the eigenvalues;

Ratio We selected the eigenvalues whose ratio with the highest eigenvalue was above a given threshold.

For computational complexity reasons, we also limited to 200 the maximum rank of the quantum density.

2. Following (Piwowarski et al., 2010), the topical space was approximated by the term space where each dimension corresponds to uni-grams, bi-grams, or either. Further, in a term space, various weighting schemes (e.g. TF-IDF) exist, and we select for each model the best performing one in Section 5.4;
3. Choosing the mixture weights (as defined in Section 4.1).

To avoid over-fitting, we chose the parameters using two DUC collections (e.g., 2005 and 2007), evaluating on the held-out one (e.g., 2006) *only at the end of the three steps*. The evaluation performed on the held-out collection is presented in Section 5.6.

At each step, and for each model, to select a set of parameters among P_1, \dots, P_p , we proceed as follows. For each parameter set P_i , we performed a paired one-sided t -test on the difference of performance (for both the ROUGE-2 and ROUGE-SU4 metrics) with all the P_j ($i \neq j$) to check whether P_i performed worse than P_j . We then computed the minimum p_i of the p -values of the t -tests for all $j \neq i$. The value p_i represents the minimum probability to wrongly discard P_i in favour of another set of parameters. The selected sets are those for which the probability p_i of wrongly discarding are at least half of the highest of these probabilities, that is those for which $p_i/p \geq 0.5$ where $p = \max_i p_i$. For example if, for a given set of parameters, the maximum p -value is 0.7, then the probability of being wrong by selecting another approach would be 0.7 (this number was chosen empirically on preliminary experiments, and does only select a few, typically one, system), and we would select all the set of parameters such that the minimum probability of being wrong is over 0.35. At the end of the last step, to select only one system for each model, we selected the parameters with respect to the ROUGE-2 metric and chose only the one with the highest minimum p -value.

Note that for each of the systems, a summary is formed by first ordering sentences with respect to their scores (e.g. quantum probabilities). We take the highest scored sentence as the lead and add other high scored sentences to the summary using a Traveling Salesman (TS) formulation (Reinelt, 1994). This selection is done in two steps; first, we compute a similarity measure, t_{ij} , between some pairs of sentences (s_i, s_j) in the top 15 scored sentences

$$\forall (s_i, s_j) \in \mathcal{T}_{15}; t_{ij} = 1 - \frac{n_{ij}}{\sqrt{n_{ii}n_{jj}}}$$

where, n_{ij} is the number of common terms in s_i and s_j . For sentences in the same document this number is doubled. In the second step, we determine an ordering that minimises the sum of the similarities between adjacent sentences. Sentences are added with the final summary length constraint of 250 words. This selection technique was used by one of the best performing systems at DUC 2006 (Conroy, Schlesinger, O’Leary & Goldstein, 2006).

5.3 Rank selection

The first series of experiments investigated the effect of rank selection on the different approaches. We experimented with the three different selection strategies described in Section 5.2. In the case of density, for the *ratio* strategy, we experimented with values from 0.2 to 0.8 by steps of 0.1. The corresponding rank values are shown in Figure 4 (note that rank was limited to 200 for computational reasons - which is reasonable given that most of the chosen ranks are already below this limit). For the *maximum* strategy, we used the values 1, 5, 10, 25, 50 and 100.

Figures 5 and 6 report the average difference between the given settings and the mean performance for a topic over all the model and parameter settings, for the maximum (left) and ratio (right) rank selection strategies. Summary of values are reported through boxplots thus showing five important pieces of information namely the minimum, first, second (median), third, and maximum quartiles. Overall, we observe that rank reduction is beneficial, since high ranks (Ratio 0.2 or Max 100) do not perform well whatever the model. We also notice that the QIA-based approach perform better in median, whatever the parameter settings. We can then distinguish three different behaviours depending on the model.

First, Steinberger and Ozsoy models are those for which rank selection has the most important effect. In particular, low ranks do not perform well and high ranks are even worse. This is sensible since with low ranks only sentences corresponding to the same atomic topics are selected, whereas with high ranks all sentences scores are close to 1, and thus selection has more to do with random noise than with the topicality of the sentences.

Second, for Gong and Murray models, we can see that low rank selection is not a good strategy, for the same reasons as for Steinberger and Ozsoy. Gong’s model performance also decreases with high ranks, but Murray’s is not affected by this, which is normal since Murray modified Gong algorithm so that more important sentences (higher eigenvalues) are selected more often within the first extracted sentences: Hence, sentences associated with atomic topics whose eigenvalue is low are selected much latter.

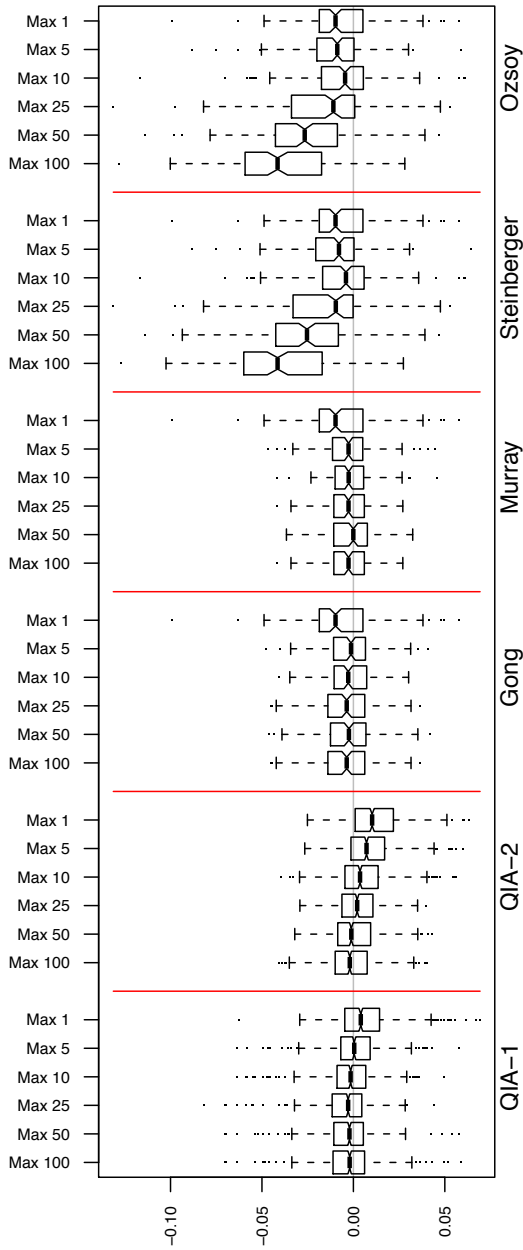


Figure 5: Boxplots of the difference with the mean value of each topic for the ROUGE-2 metric for the maximum rank selection method

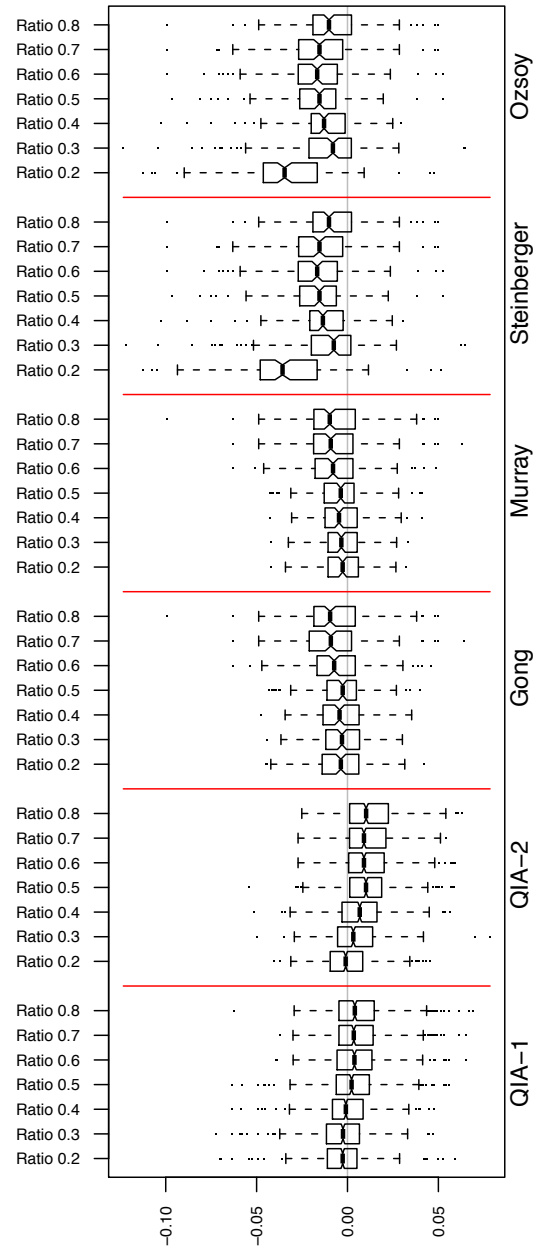


Figure 6: Boxplots of the difference with the mean value of each topic for the ROUGE-2 metric for the ratio rank selection method

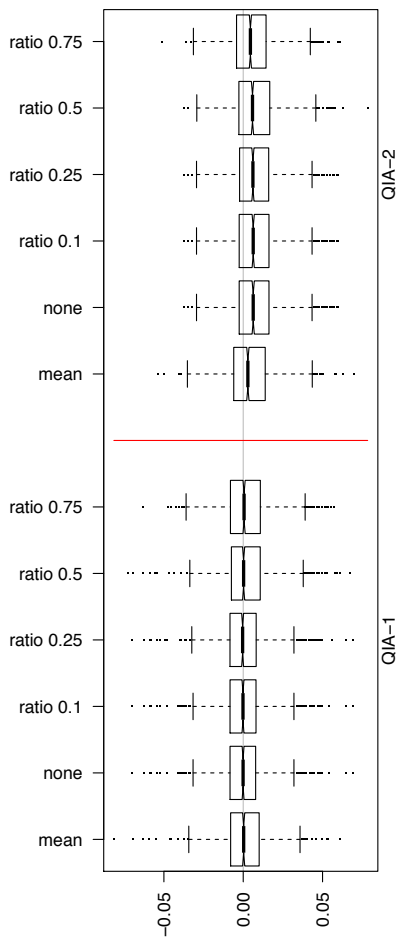


Figure 7: Boxplots of the difference with the mean value of each topic for the ROUGE-2 metric for the different subspace rank selection strategies, for the QIA-based models

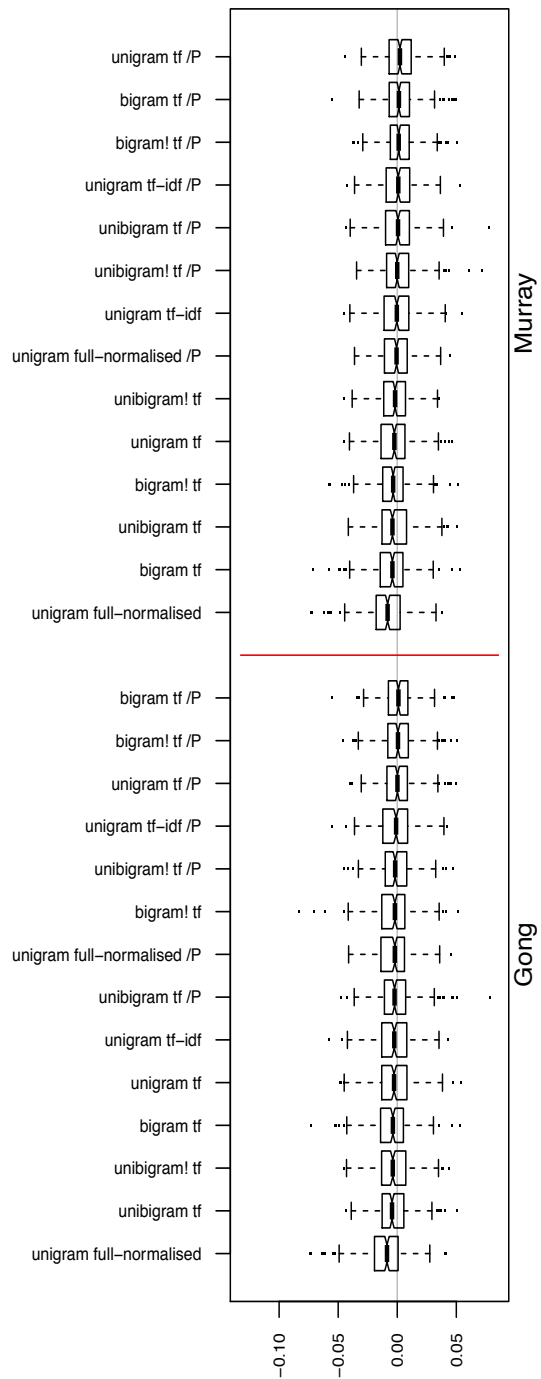


Figure 8: Gong and Murray – Boxplots of the difference with the mean value of each topic for the ROUGE-2 metric for different sentence representation schemes (strict bigram is indicated by a “!”)

Finally, the two QIA-based approaches work in general better with low ranks, e.g. close to the minimum 1, which would indicate that in most cases, there is just one main topic to be summarised, and the sentences should be selected so as to cover as much of it.

For the QIA-based approaches, we are also interested by the subspace rank selection. Results are reported in Figure 7. We observe that the QIA models are not affected much by the subspace rank selection – given the variance of the results, the subspace rank should hence be chosen depending on the other parameters. As shown in the final evaluation (Section 5.6), the rank selection that was chosen for each QIA model tend to preserve most of the dimensions of the subspace (ratio ≥ 0.5), which means that it is better to preserve the full subspace covered by all the sentences of the extracted summary.

5.4 Sentence representation

In the second set of experiments, we looked at the representation of sentences, i.e. by changing the vector space to which vectors representing the sentences belong. More precisely, we varied the indexed units and the weighting scheme.

For the indexed units, we used unigrams, and, motivated by the bi-gram based summarisation system that performed best at DUC 2006 (Jagarlamudi, Pingali & Varma, 2006), we also experimented with bi-grams, and a combination of both (i.e. both uni-grams and bi-grams). For bi-grams, we either selected bi-gram of words separated by stop words or not (strict bi-grams). The former was used because intuitively, bi-grams are usually important if their constituents are close together. Following the findings of Ozsoy in LSA-based summarisation (Ozsoy et al., 2010), we also experimented by restricting the indexed units to be nouns or named entities (categories NN and NP) using a part-of-speech tagger (Schmid, 1994). Results are reported in Figures 8 (Gong and Murray), 9 (Steinberger and Ozsoy) and 10 (QIA) as boxplots of the difference between the evaluated system and the mean performance value of all systems for the ROUGE-2 metric since the ROUGE-SU4 metric showed the same pattern of performance.

In the case of TF-based approaches and unigram index terms (or uni- and bi-gram index terms), restricting to noun part-of-speech was beneficial to all systems, which matches the conclusions drawn in (Ozsoy et al., 2010), and is intuitive since this filters out many of the non necessary information when building up summaries.

Among the weighting schemes, TF (for LSA-based models) and TF-IDF (for QIA) worked the best. TF has been reportedly a good performing weighting scheme for extractive summarisation, since the IDF information is not that important within a set of topic-biased documents, so it is interesting that using IDF information works better for the two QIA-based models.

A key difference, especially with low-rank densities, between the QIA and LSA-based models is the fact that with the QIA models a subspace is built that corresponds to the constructed summary. Using a TF scheme can dramatically change the shape of this subspace. Indeed, consider for example the pseudo-sentences, s_1 = “the sentence”, s_2 = “the paragraph” and s_3 = “a paragraph”. With a TF-IDF approach, the subspace corresponding to $\{s_1, s_2\}$ would be very close to the subspace $\{s_1, s_3\}$ whereas it would not be the case with the TF weighting scheme.

One way to verify the above hypothesis about the importance of the defined subspaces in QIA is to look at the difference of performance when using part-of-speech (POS) filtering with the TF or TF-IDF weighting

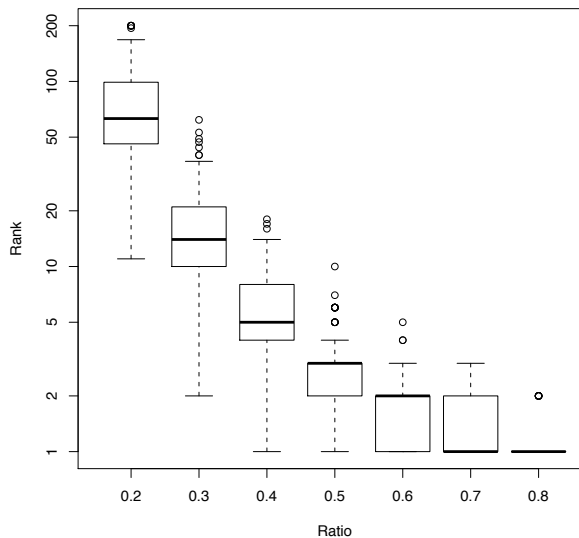


Figure 4: Boxplots of the final rank for the different ratio selections

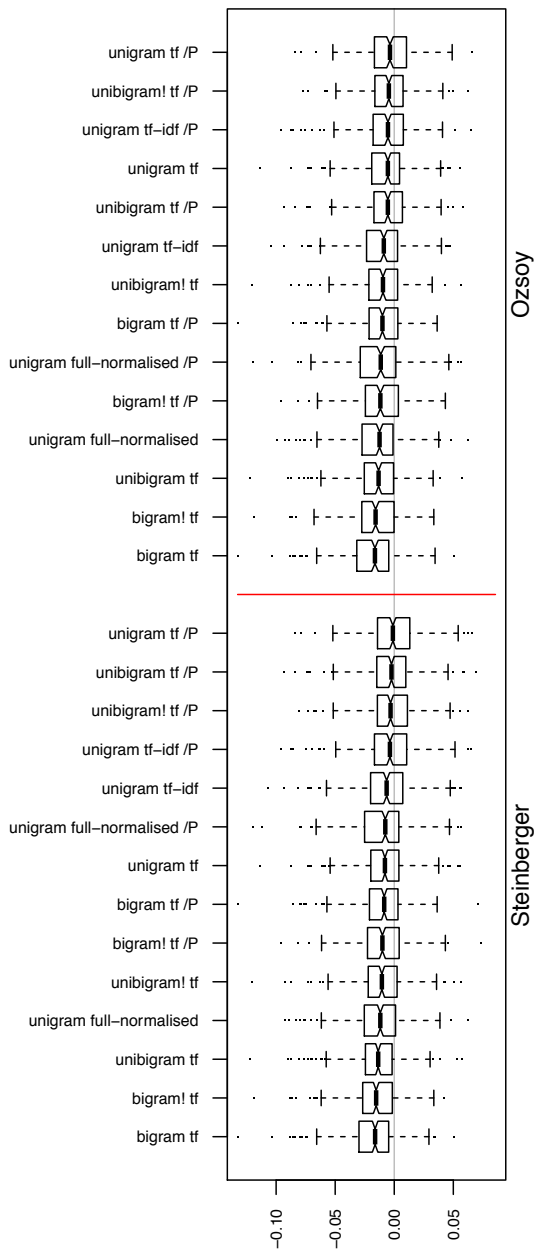


Figure 9: Steinberger and Ozsoy – Boxplots of the difference with the mean value of each topic for the ROUGE-2 metric for different sentence representation schemes (strict bigram is indicated by a “!”)

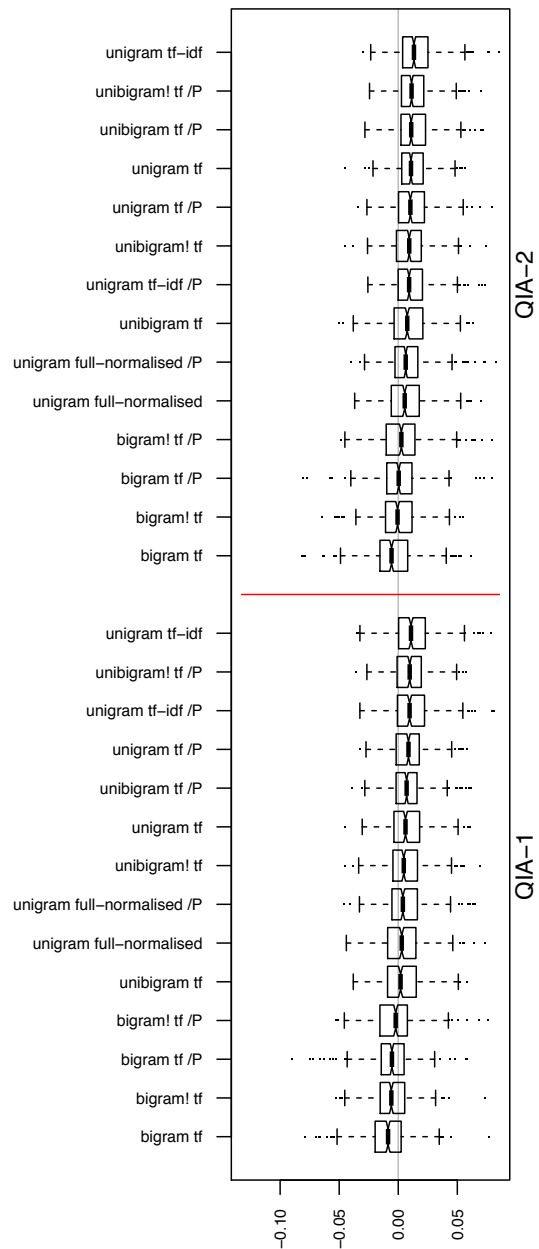


Figure 10: QIA models – Boxplots of the difference with the mean value of each topic for the ROUGE-2 metric for different sentence representation schemes (strict bigram is indicated by a “!”)

schemes. Since POS filters out more often units with low IDF, the difference in performance should be of greater magnitude with TF, and this corresponds to what we observe in the results (with the ROUGE-2 metric, the mean absolute difference between QIA and LSA approaches is of 0.17 for TF versus 0.15 for TF-IDF).

As a final note, the representation of sentences is an open topic in the QIA framework. The quantum formalism provides the possibility to use the complex field, instead of the real field as followed in this paper, hence offering another degree of freedom of the representation since each component of the vector could be a complex number. It is however difficult to know at this point how the complex field can be leveraged, and we refer the readers to (Zuccon, Piwowarski & Azzopardi, 2011) for a short discussion on this topic.

5.5 Prior sentence distribution

In the third set of experiments, we investigated the weight in the mixture defined in Equation 10, that is with α_0 (uniform), α_d (document), α_l (length) and α_t (topic). We varied the values of each parameter within the set 0, 0.25, 0.50 and 1, ensuring that weights were summing up to 1.

We first run an ANOVA on each model to look at the effect of each parameter. The results did not vary depending on the model. The parameters that had the most important effect are document, topic and length priors (in order of significance). We found a significative interaction between topic on the one hand, and document or length on the second hand, which in practice means that if we set the topic prior, then document and length prior influence the performance independently.

Results are reported in Figures 11 (Gong and Murray), 12 (Steinberger and Ozsoy) and 13 (QIA) as boxplots of the difference between the evaluated system and the mean performance value of all systems for the ROUGE-2 metric.

First, we can see that LSA-based approaches were more affected by the change in mixture weights than the QIA-based ones, so it is an important parameter for these approaches only – hence most of the conclusions here apply to LSA-based models.

When we look at the different classes of models, we can distinguish three more detailed effects of the priors:

Gong/Murray (LSA-I) performs better with topic prior ($\alpha_t = 0.25$) and length prior ($\alpha_d = 0.25$) but no document prior;

Steinberger/Ozsoy (LSA-II) performs better with topic prior ($\alpha_t = 0.25$) and document prior ($\alpha_d = 0.25$) but no length prior;

QIA performs better with no topic prior. In particular, the uniform prior did perform well for both QIA-1 and QIA-2.

From these observations, we can state the following. First, all LSA-based approaches need to have some weight on sentences containing topic terms. Second, QIA-based models are able to implicitly capture the topic at hand from the documents provided for summarisation and are less sensitive to varying documents or sentence lengths. This shows that when the documents to be summarised are on-topic, as it is the case in the DUC test collections, there is no need for the QIA approaches to use any information about the topic that was used to select those documents.

5.6 Evaluations on the held-out collection

The last set of experiments was conducted to evaluate the optimised models (i.e. the different models where the parameters were selected as described in Section 5.2) on the held-out DUC corpus. Thus, the results reflect the results we would have obtained on one DUC collection, when the summaries of the two others are available for parameter tuning.

We also compared the results with two graph-based models (symmetric non-negative matrix factorisation (SNMF) and Lexrank); two baseline systems, namely `lead` and `random`; and the best competing summarisation system in DUC 2005, DUC 2006 and DUC 2007, denoted by Best@DUC.

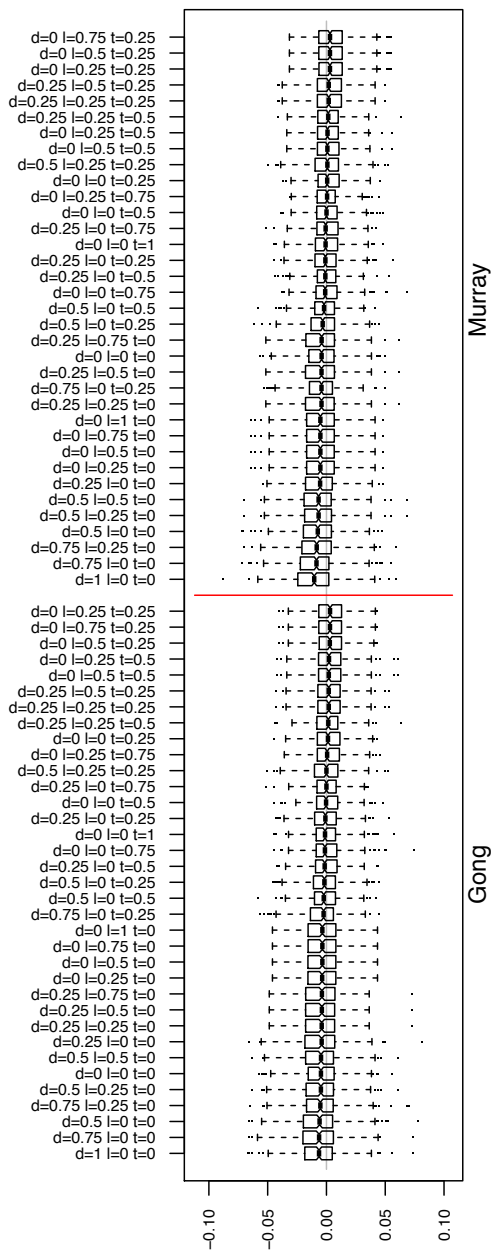


Figure 11: Boxplots of the difference with the mean value of each topic for the ROUGE-2 metric for different mixture settings. The letters d, l and t are respectively for document, length and topic bias weights.

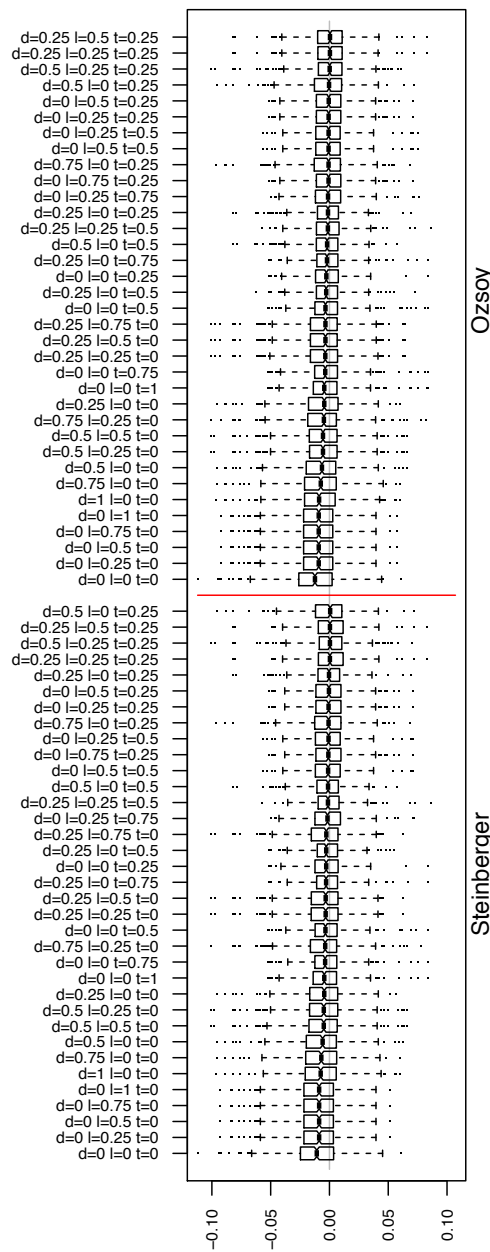


Figure 12: Boxplots of the difference with the mean value of each topic for the ROUGE-2 metric for different mixture settings. The letters d, l and t are respectively for document, length and topic bias weights.

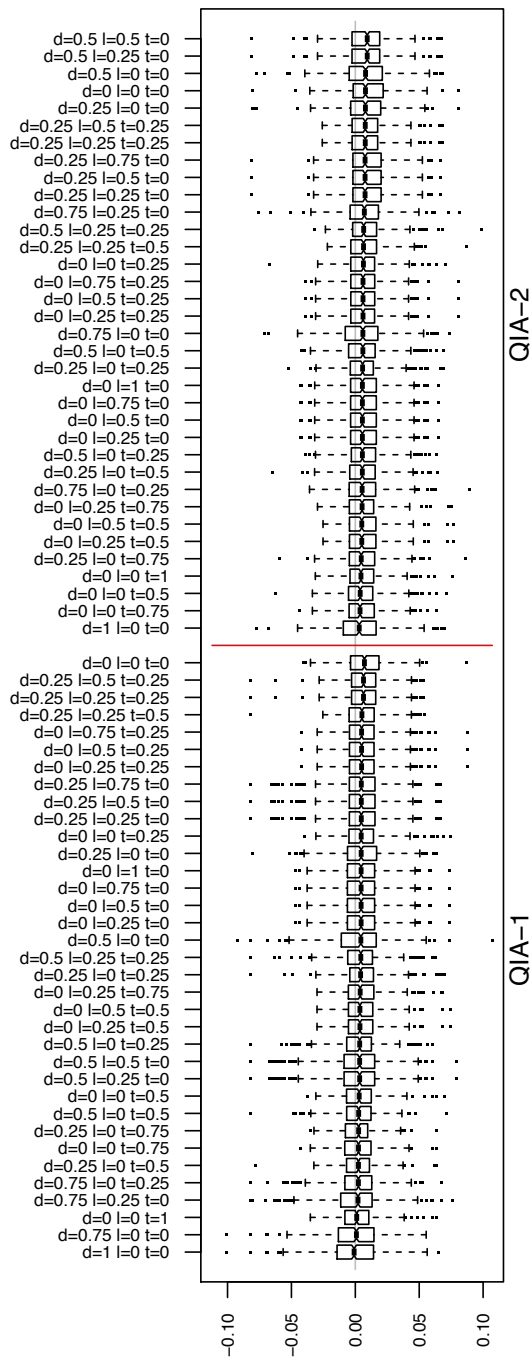


Figure 13: Boxplots of the difference with the mean value of each topic for the ROUGE-2 metric for different mixture settings. The letters d, l and t are respectively for document, length and topic bias weights.

Metric	DUC 2005		DUC 2006		DUC 2007	
Model / DUC	ROUGE-2	ROUGE-SU4	ROUGE-2	ROUGE-SU4	ROUGE-2	ROUGE-SU4
Best@DUC	0.072	0.133	0.095	0.155	0.123	0.175
Average@DUC	0.060	0.115	0.075	0.132	0.096	0.150
Lead	0.043	0.093	0.053	0.104	0.065	0.113
Random	0.041	0.091	0.049	0.101	0.060	0.110
LexRank	0.076	0.136	0.093	0.150	0.120	0.172
SNMF	0.060	0.121	0.085	0.140	0.110	0.158
Gong ^a	0.057	0.112	0.076	0.136	0.100	0.155
Murray ^a	0.056	0.109	0.076	0.135	0.104	0.159
Ozsoy ^a	0.050	0.099	0.072	0.128	0.090	0.140
Steinberger ^a	0.050	0.099	0.072	0.128	0.089	0.140
QIA-1 ^a	0.062	0.117	0.089	0.147	0.123	0.181
QIA-2 ^a	0.068	0.124	0.093	0.151	0.116	0.175
Gong ^b	0.062	0.121	0.083	0.143	0.112	0.171
Murray ^b	0.063	0.122	0.083	0.143	0.113	0.172
Ozsoy ^b	0.077	0.137	0.072	0.128	0.092	0.146
Steinberger ^b	0.077	0.137	0.081	0.143	0.091	0.146
QIA-1 ^b	0.077	0.135	0.091	0.152	0.127	0.185
QIA-2 ^b	0.080	0.141	0.097	0.159	0.118	0.179
Gong ^c	0.072	0.133	0.087	0.148	0.118	0.180
Murray ^c	0.073	0.135	0.086	0.147	0.120	0.181
Ozsoy ^c	0.071	0.133	0.085	0.145	0.111	0.173
Steinberger ^c	0.071	0.133	0.081	0.144	0.111	0.169
QIA-1 ^c	0.077	0.135	0.091	0.151	0.127	0.185
QIA-2 ^c	0.080	0.141	0.097	0.159	0.125	0.183

Table 3: Final evaluation on the held-out corpus. The first three rows give respectively the performance of the best system in DUC, the random and lead strategies. There are three series of results for each LSA and QIA based approaches: (a) after the rank selection of Section 5.3, (b) after the weighting scheme selection of Section 5.4, and (c) after the mixture weights selection of Section 5.5. Best performances are indicated by boldface.

		Density	Subspace	Weighting	Indexed unit	POS	α_0	α_d	α_l	α_t
Gong	2005	Max 50		TF	strict bigram	y	0.25		0.50	0.25
	2006	Max 5		TF	unigram	y	0.50		0.25	0.25
	2007	Max 50		TF	strict bigram	y	0.25		0.50	0.25
Murray	2005	Max 50		TF	strict bigram	y	0.50		0.25	0.25
	2006	Max 10		TF	unigram	y	0.50		0.25	0.25
	2007	Max 50		TF	strict bigram	y	0.50		0.25	0.25
Ozsoy	2005	Max 1		TF	unigram	y	0.75		0.25	
	2006	Max 10		TF	unigram	y	0.25	0.25	0.25	0.25
	2007	Max 10		TF	strict bigram	y	0.25		0.50	0.25
Steinberger	2005	Max 1	TF	unigram	y	0.75		0.25		
	2006	Max 1	TF	unigram	y	1.00				
	2007	Max 10	TF	strict bigram	y	0.25	0.25	0.25	0.25	
QIA-1	2005	Max 1	Ratio 0.75	TF-IDF	unigram	n	1.00			
	2006	Ratio 0.8	Mean	TF-IDF	unigram	n	0.25	0.25	0.25	0.25
	2007	Ratio 0.8	Ratio 0.75	TF-IDF	unigram	n	1.00			
QIA-2	2005	Max 1	None	TF-IDF	unigram	n	1.00			
	2006	Ratio 0.8	Ratio 0.25	TF-IDF	unigram	n	0.5	0.5		
	2007	Ratio 0.5	Ratio 0.25	TF	unigram	n		0.5	0.25	0.25

Table 4: Parameters for the different models whose performance is shown in Table 3

The lead baseline returns all the first sentences (up to 250 words) in the most recent document for each topic and the random baseline selects sentences randomly.

SNMF conducts symmetric non-negative matrix factorisation on a sentence-sentence similarity matrix (Wang et al., 2008), the hyper-parameter λ for computing sentence scores was fixed to 0.7 which gave best results on all three DUC collections.

Lexrank defines a random walk model on top of a graph where sentences to be summarised define its nodes and the edges represent the similarity measures between the nodes of the graph. Sentences are then scored by the expected probability of a random walker visiting each sentence (Erkan & Radev, 2004). Here, the cosine threshold t was fixed to 0.1 leading to best results with this approach.

The selected systems for each model are reported in Table 4, the corresponding results in Table 3 and the pairwise t-tests in Table 5. We can see that our results match the main conclusion drawn in the previous sections, although parameters vary slightly depending on the specific corpora on which they were optimised. More precisely, for LSA-based approaches, TF and unigram/strict bi-grams with POS filtering perform the best, and including the different priors was important. The parameters are quite different for QIA-based models, where a TF-IDF weighting scheme on unigrams, with uniform prior over sentences, perform the best in general.

From a performance point of view, we improved substantially all the LSA-based models by selecting appropriate indexing units (in particular, using part-of-speech tagging, as suggested in (Ozsoy et al., 2010)) and using priors on sentences in the document to be summarised, as suggested by the QIA approach. Those priors (as shown in Table 4) are biased towards the topic and the length of the sentences.

The QIA-2 model is slightly superior to the QIA-1 except on DUC 2007. This shows that the QIA main hypothesis, that any linear combination of atomic topics present in a document is also a topic of the document, as discussed in Section 4.2, does hold in the case of summarisation, as the QIA-1 approach performed well in comparison with QIA-2.

In all cases, we can observe that the QIA-based models perform the best for both metrics. The performance of both QIA-based models are over those of the best systems in DUC for the corresponding years (not significant

	b	l	s	G	M	O	S	Q1	Q2	b	l	s	G	M	O	S	Q1	Q2	
	ROUGE-2									ROUGE-SU4									
2005																			
best@DUC			***	+		+	+					***							
LexRank	***		***	*	+	+	+			***		***	*	+	+	+	+		
SNMF																			
Gong			***			+	+			+		***			+	+			
Murray	+		***	+		+	+			+		***	+		+	+			
Ozsoy			***							+		***							
Steinberger			***							+		***							
QIA-1	+	+	***	+	+	*	*			+		***	+	+	+	+			
QIA-2	**	+	***	**	**	**	**	*		**	+	***	*	*	*	*	**		
2006																			
best@DUC		***	**	*	**	***	**	+			***	**	*	**	***	**	+		
LexRank			*	*	*	*	**	+				***	+	+	*	*			
SNMF						+	+												
Gong			+		+	+	*					**		+	+	+			
Murray			+			+	+					**			+	+			
Ozsoy							+					**				+			
Stein.												+							
QIA-1			+	+	*	***	***				+	***	+	**	**	**			
QIA-2	+	+	+	**	***	***	*	*		+	+	**	+	+	*	*	+		
2007																			
best@DUC		***	**	+	+	**	**				***	***			+	+			
LexRank			*	+	+	*	*					**			+	+			
SNMF																			
Gong			*			+	+			+	*	***			*	*			
Murray			*	+		*	*			*	**	***	+		*	*			
Ozsoy			+									*							
Stein.			+									*			+				
QIA-1	+	*	**	**	*	***	***	+		***	***	***	*	+	**	**		+	
QIA-2	+	*	**	*	*	**	**			***	***	***	+	+	**	**			

Table 5: Pairwise t -tests between all selected systems. The “+” sign means that system (row) performed better than another (column), but not significantly. The number of stars varies between 1 and 3 and corresponds to significance levels of respectively 0.05, 0.01 and 0.001. For space reasons, we use only one letter in the columns to denote the different systems (the order is the same as for the rows).

except for ROUGE-SU4 in DUC 2005 and 2007); in particular, this means that in 2007 QIA-based models would have been ranked first since the data from 2005 and 2006 was available.

Finally, in all cases, QIA-based models are in most of the cases performing better (significantly in 2007 for Lexrank and 2005-07 for SNMF) than two state of the art extractive summarisation methods, namely SNMF and Lexrank, thus showing that the QIA framework is a very promising approach for extractive summarisation.

5.7 Summary

In summary, our experimental results show that when summarisation is performed on a set of relevant documents to a given topic (topic-oriented documents), as it is the case with the DUC collections, QIA-based models are able to implicitly capture the topics covered by the set of documents and are less sensitive to varying documents or sentence lengths. This is an important result as it means that the similarity estimations between sentences and the topic, performed by most systems in these competitions, is not required by the QIA-based models. Indeed, the latter uncover automatically, without relying explicitly on the DUC-provided topic at hand, the important atomic topics covered by a set of topic-oriented documents

More precisely, we showed that even though LSA and QIA-based techniques are based on spectral decomposition, these models differ in the choice of their optimal parameters. LSA-based approaches benefit from the various pre-processing steps (part-of-speech, bi-grams, topic and length bias, rank selection) whereas QIA-based approaches rely on the standard IR TF-IDF scheme and a few (typically one) atomic topics that represent the important topics of the documents to summarise. This difference is due to the criteria used to select sentences. LSA-based models do not consider the “topical” space covered by a set of extracted sentences, whereas the from QIA-based models do.

This leads to an important conclusion. The topical space, in the case of summarisation, resembles more a TF-IDF term space than a TF term space, which can be linked to the QIA hypothesis on the linear combination of atomic topics. Such a linear combination makes more sense when less important terms (i.e. low IDF) do not influence much the result of the linear combination.

Finally, we showed that QIA-based models performed better (significantly in one DUC collection) than the best systems that competed in the DUC competitions and than to state of the art extractive models, namely SNMF and Lexrank. For illustration purposes, in Appendix A we provide an example of the summary extracted by the different systems where the QIA framework is shown to correctly identify and extract sentences corresponding to the most important topics.

6 Conclusion

In this paper, we described an approach for multi-document summarisation (MDS) motivated by the Quantum Information Access (QIA) framework, which in turn is based on the quantum probability theory. The results we found are of great importance, both from a theoretical and practical point of view.

From a theoretical point of view, we showed that it was possible to interpret in a principled and (quantum) probabilistic way the successful Singular Value Decomposition (LSA) approaches to summarisation and, more interestingly, to identify their limitations from a purely quantum probabilistic theoretic interpretation. So far, only intuitive arguments were put forward.

This theoretic analysis brought two important results. First, we showed that it is possible to modify LSA-based approaches so that they benefit from the QIA framework, leading in practice to a much improved performance in the DUC collection with respect to the most important metrics (ROUGE-2 and ROUGE-SU4). Second, an analysis of the limitations of LSA-based approaches provided a new and more natural QIA-based criterium to build summaries. This criterium provides a global measure of the quality of the summary by measuring to which extent the topics of the selected sentences “cover” the important topics of the documents to be summarised. This is to be contrasted to the LSA-based criteria, which consider sentences in isolation.

Extensive experiments show that the theoretic insights translate into a difference in performance. The QIA-based approach not only perform much better than previous LSA approaches, but is also competitive

with state-of-the-art summarisation approaches (including two state of the art models, SNMF and Lexrank). Indeed, it performed better than the best performing systems in DUC 2005, DUC 2006 and DUC 2007.

Another finding is about the validation of one of the fundamental hypothesis of the QIA framework, which states that if two atomic topics are present in a document, then any linear combination of these atomic topics in the topical vector space is also an atomic topic of the document. This hypothesis has not been verified so far within the QIA framework. This is bringing us new insights into to the potential of quantum theory and on the importance of choosing the right representation, i.e. the topical space in this paper, as well as a new momentum to explore the application of the QIA framework in Information Access and related areas.

It is our belief that the potential of the QIA goes beyond what we presented in this paper. A main extension to this work is to use kernels, that have been useful in many machine learning algorithms relying on inner products like Support Vector Machines (Schölkopf & Smola, 2002). Indeed, these, by defining how to compute an inner product in a vector space without explicitly computing the vectors, allows to work in higher (possibly infinite) dimensional Hilbert spaces. This for instance would allow to work in spaces more complex than uni/bi-grams term spaces, and to integrate semantic and syntactic information for summarisation purpose, thus exploring further the question of what the topical space should look like. An interesting possibilities would be to provide a mean to build sentence bi-gram models, thus addressing a long standing problem in text summarisation – how to select the sentence that is the most likely to follow another one in the summary. This is part of our future work.

As a final remark, this is the first time that the QIA framework is being used for other tasks than ad-hoc IR, e.g. (Piwowarski et al., 2010), and hence shows the potential of QIA for Information Access tasks, and more generally of using the quantum probability theory outside physics: Without the quantum formalism, and the link between geometry as used in IR and this formalism advocated by van Rijsbergen (Rijsbergen, 2004) and Widdows (Widdows, 2004) on one hand, and the QIA framework methodology on the other hand, it would have been impossible to give a quantum probabilistic interpretation of previous LSA-based approaches and propose a new and better criterion for sentence selection in extractive summarisation.

7 Acknowledgements

This research was supported by an Engineering and Physical Sciences Research Council grant (Grant Number EP/F015984/2) and by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement N. 247590.

References

- Amini, M. R. & Usunier, N. (2011). Transductive learning over automatically detected themes for multi-document summarization. In *Proceedings of the 34th annual international ACM SIGIR conference* (pp. 1193–1194).
- Amitay, E. (2001). Trends, fashions, patterns, norms, conventions . . . and hypertext too. *Journal of the American Society for Information Science and Technology*, 52(1), 36–43.
- Barzilay, R. & Elhadad, M. (1997). Using lexical chains for text summarization. In *In proceedings of the acl workshop on intelligent scalable text summarization* (pp. 10–17).
- Barzilay, R., McKeown, K. R. & Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *Proceedings of the 37th ACL conference* (pp. 550–557).
- Chen, Y., Wang, X. & Liu, B. (2005). Multi-document summarization based on lexical chains. In *Proceedings of 2005 international conference on machine learning and cybernetics* (pp. 1937–1942).
- Conroy, J. M., Schlesinger, J. D., O’Leary, D. P. & Goldstein, J. (2006). Back to basics: Classy 2006. In *DUC-NIST Proceedings Document Understanding Conference (DUC)*.
- Deerwester, S., Dumais, S., Furnas, G. & Landauer, T. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science and Technology*, 41(6), 391–407.

- Erkan, G. & Radev, D. R. (2004). Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22, 457-479.
- Fang, H., Tao, T. & Zhai, C. (2011, April). Diagnostic evaluation of information retrieval models. *ACM Transactions on Information Systems*, 29, 7:1-7:42.
- Gong, Y. & Lin, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference* (pp. 19-25).
- Halteren, H. V. & Teufel, S. (2003). Examining the consensus between human summaries: initial experiments with factoid analysis. In *HLT/NAACL-2003 Workshop on Automatic Summarization*.
- Harabagiu, S. & Lacatusu, F. (2005). Topic themes for multi-document summarization. In *Proceedings of the 28th annual international ACM SIGIR conference* (pp. 202-209).
- Hirao, T., Sasaki, Y. & Isozaki, H. (2001). An extrinsic evaluation for question-biased text summarization on QA tasks. In *Proceedings of the ANLP/NAACL Workshop on Automatic Summarization* (pp. 61-68).
- Huertas-Rosero, A. F., Azzopardi, L. A. & Rijsbergen, C. J. (2009). Eraser lattices and semantic contents. In *Proceedings of the 3rd international symposium on quantum interaction* (pp. 266-275).
- Jagarlamudi, J., Pingali, P. & Varma, V. (2006). Query independent sentence scoring approach to DUC 2006. In *Proceedings of the Document Understanding Conference*.
- Knight, K. & Marcu, D. (2002). Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence*, 139, 91-107.
- Li, J. & Sun, L. (2008). A lexical chain approach for update-style query-focused multi-document summarization. In *Proceedings of the 4th Asia Information Retrieval Conference on Information Retrieval Technology* (pp. 310-320).
- Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL'04 Workshop* (pp. 74-81). Association for Computational Linguistics.
- Lin, C. Y. & Hovy, E. (2002). From single to multi-document summarization: A prototype system and its evaluation. In *Proceedings of 40th ACL conference* (pp. 457-464).
- Lin, C. Y. & Hovy, E. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* (pp. 71-78).
- Mani, I. & Bloedorn, E. (1999). Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1-2), 35-67.
- McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R. & Eskin, E. (1999). Towards multidocument summarization by reformulation: progress and prospects. In *Proceedings of the 16th conference on AAAI/IAAI* (pp. 453-460).
- McKeown, K. R., Passonneau, R. J., Elson, D. K., Nenkova, A. & Hirschberg, J. (2005). Do summaries help? In *Proceedings of the 28th annual international ACM SIGIR conference* (pp. 210-217).
- Melucci, M. (2008). A basis for information retrieval in context. *ACM Transactions on Information and System*, 26, 1-41.
- Mihalcea, R. (2005). Language independent extractive summarization. In *Proceedings of the ACL 2005 conference* (pp. 49-52).
- Mitra, M., Singhal, A. & Buckley, C. (1997). Automatic text summarization by paragraph extraction. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization* (pp. 39-49).
- Murray, G., Renals, S. & Carletta, J. (2005). Extractive summarization of meeting recordings. In *Proceedings of the 9th european conference on speech communication and technology* (pp. 593-596).
- Nielsen, M. A. & Chuang, I. L. (2000). *Quantum computation and quantum information*. New York, NY, USA: Cambridge University Press.
- Ozsoy, M. G., Cicekli, I. & Alpaslan, F. N. (2010). Text summarization of turkish texts using latent semantic analysis. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 869-876).
- Piwovarski, B., Frommholz, I., Lalmas, M. & Rijsbergen, K. van. (2010). What can quantum theory bring

- to information retrieval? In *Proceedings of the 19th ACM Conference on Information and Knowledge Management* (pp. 59–68).
- Piwowarski, B., Trotman, A. & Lalmas, M. (2009, jan). Sound and complete relevance assessments for XML retrieval [Journal]. *ACM Transactions On Information Systems*, 27(1).
- Radev, D. R., Jing, H., Styś, M. & Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing Management*, 40, 919–938.
- Reinelt, G. (1994). *The traveling salesman: computational solutions for tsp applications*. Berlin, Heidelberg: Springer-Verlag.
- Rijsbergen, C. J. van. (2004). *The geometry of information retrieval*. Cambridge University Press.
- Sampath, G. & Martinovic, M. (2002). A multilevel text processing model of newsgroup dynamics. In *Proceedings of international conference on applications of natural language to information systems* (pp. 208–212).
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*.
- Schölkopf, B. & Smola, A. (2002). *Learning with kernels*. MIT press.
- Sparck Jones, K. (1993). *Discourse modeling for automatic summarizing* (Tech. Rep.). Computer Science Department, University of Cambridge.
- Steinberger, J. & Ježek, K. (2004). Using latent semantic analysis in text summarization and summary evaluation. In *Proceedings of the Information System Implementation and Modeling conference*.
- Turpin, A., Tsegay, Y., Hawking, D. & Williams, H. E. (2007). Fast generation of result snippets in web search. In *Proceedings of the 30th annual international ACM SIGIR conference* (p. 127-134).
- Wang, D., Li, T., Zhu, S. & Ding, C. (2008). Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the 31th annual international ACM SIGIR conference* (pp. 307–314).
- Widdows, D. (2004). *Geometry and meaning*. University of Chicago Press.
- Zuccon, G. & Azzopardi, L. (2010). Using the quantum probability ranking principle to rank interdependent documents. In *Proceedings of the 32nd European Conference on Information Retrieval* (p. 357-369).
- Zuccon, G., Azzopardi, L. A. & Rijsbergen, C. J. (2009). Semantic spaces: Measuring the distance between different subspaces. In *Proceedings of the 3rd international symposium on quantum interaction* (pp. 225–236).
- Zuccon, G., Piwowarski, B. & Azzopardi, L. (2011). On the use of complex numbers in quantum models for information retrieval. In *Advances in Information Retrieval Theory: Third International Conference*.

A Summary extracts (topic 385 - DUC 2005)

Topic 385 - *What is the current status of research and development on electric automobiles? What are the positive and negative factors for their usage? Which companies are involved in their development?*

Apart from the human summary, all the summaries are extractive. They were selected as an example of when QIA-1/2 performs better than all other methods. The human summary is provided only for reference.

In the human summary we can identify the following topics regarding electric cars: future of cars, the efforts, the legislation, and the ecological advantages and difficulty to build electric cars. We can make the following observations:

- LexRank and SNMF both fail to identify several topics: future of the car (SNMF), legislation (both), difficulty (LexRank);
- Gong and Murray miss some topics, probably because of the hard clustering that characterises those methods (e.g., the difficulty of making electric cars);
- Ozsoy and Steinberger suffer from the same problem of sampling again and again the same topics (efforts);
- The QIA approach succeeds in extracting sentences covering the major topics.

A.1 Human summary

Huge research efforts in viable electric cars has been going on the past several years. Carmakers around the world see electric vehicles as the only available technology to provide immediate pollution-free driving. A sense of urgency was prompted by the California legislature's calling for 2% of car manufacturers' sales to be of "zero-emission vehicles" from 1998, rising to ten percent by the 21st century. Up to twelve other states are seriously considering adopting similar requirements. The electric car is considered by many to be cheap to run, virtually silent, non-polluting, and easy to drive, providing good acceleration and reasonable highway cruising. Nearly all major car manufacturers – GM, Ford, Chrysler, Daimler-Benz, Renault, Peugeot, Ford, VW, and BMW, have made battery-powered conversions of their smaller gasoline-powered cars and delivery vehicles and are preparing plants for increased production. GM unveiled its prototype of a futuristic electric car developed from the Impact model. Severe disadvantages in manufacturing electric cars at this time are delaying extensive production. They are low on power, short of range, and expensive to make. The cost of high technology required to eliminate these problems will make them highly expensive to purchase. The biggest problem is the absence of super-efficient batteries. The industry must use cheap-lead acid batteries, which are extremely heavy and take up the rear seat space. GM, Ford, Chrysler, and federal agencies are collaborating to establish new super battery technology. For the immediate future the hybrid car, using both electricity and gasoline, is showing the most promise.

A.2 LexRank

A prototype of the electric car that BMW intends to sell in the US in the second half of the 1990s is on display at the Los Angeles motor show which opened to the public this week. Environmentalists keep on saying the battery-electric is the car of the future. Britain leads the world in the use of battery-electrics. AMERICA'S big three car makers - General Motors, Ford and Chrysler - are to co-operate much more closely in the development of electronic vehicles. Most of the weight is accounted for by the lead content. Electric cars are as old as motoring - they have been around for 100 years. This would feed current via the batteries to an electric motor driving the wheels. 'Automobiles in the future are going to be driven by fuel cells. Half the funds for the project are to be provided by the US energy department. Electric cars are the only vehicles to meet such standards so far. 'We have to look at alternatives like electric vehicles. Even the most advanced forms of battery now at the research and development stage would only improve the situation by a factor of three, according to the report. GM said it uses about a third of the energy of a conventional car. "It's consistent with the kinds of emissions standards that we are developing for the future. GM is not the only company working on electric vehicles. Ford Motor Co. is developing its own electric-powered van.

A.3 SNMF

The EV industry's development in Europe has serious implications for component suppliers. While the US automobile industry scrambles to meet 1998 deadlines to put electric vehicles on the market, controversy about the environmental benefits and commercial viability of battery-operated cars is mounting. Despite the fact that General Motors is already preparing a plant to produce the Impact, which GM intends to be the first electric car in volume production, GM, Ford and Chrysler collectively insist that other states do not have California's air pollution problems. Unlike electric vehicles that use exotic nickel-iron or sodium-sulfur batteries as power sources, the Impact uses lead-acid batteries, whose 870 pounds account for about 30% of the car's total weight. GM engineers borrowed from a prototype GM solar vehicle, the Sunraycer, to give the Impact a lightweight, aerodynamic design and improvements in motor and controls that partly account for the car's range, speed and acceleration: 0 to 60 m.p.h. in 8 seconds. A GM video showed the Impact out-accelerating Mazda Miata and Nissan 300ZX sports cars. In 1991, GM's electric vehicle programme directors implied, if not specifically stated, that cars based on the 100mph-plus, purpose-built and aluminium-bodied Impact would be rolling out of a former Buick plant at Lansing, Michigan, well before the 1998 deadline. Ken Baker, vice-president of GM's research and development centre, insists that 'GM wants electric vehicles to be a marketplace success. Americans pay little for petrol and the economic incentive towards electric cars

is zero.

A.4 Gong

The Electric Power Research Institute has worked with both GM and Chrysler to develop electric-powered vans for eventual production. AMERICA'S big three car makers - General Motors, Ford and Chrysler - are to co-operate much more closely in the development of electronic vehicles. Ken Baker, vice-president of GM's research and development centre, insists that 'GM wants electric vehicles to be a marketplace success. GM, Ford, Chrysler, electric utilities and government agencies, formed several years ago into the Advanced Battery Consortium, have awarded research and development contracts to five other battery makers pursuing alternative technologies. At the same time, the state Air Resources Board is poised to require automobile companies beginning in 1994 to begin selling a new category of low emitting vehicles that are twice as clean as the cleanest new gasoline cars on the road. Volkswagen will start production of its city car, the Chico, in 1995, and many other manufacturers have similar projects in development. It has attracted attention with the improvements it has made to a proton exchange membrane fuel cell pioneered by General Electric of the US. In that, it differs from other electrical vehicles under development, which are essentially converted delivery vans intended for commercial fleets. It required not only the development of new alloys with the required crash protection properties, but also new production processes. They will be confined mainly to city centres and could be the only kind of car allowed in the most environmentally sensitive areas.

A.5 Murray

The Electric Power Research Institute has worked with both GM and Chrysler to develop electric-powered vans for eventual production. AMERICA'S big three car makers - General Motors, Ford and Chrysler - are to co-operate much more closely in the development of electronic vehicles. Ken Baker, vice-president of GM's research and development centre, insists that 'GM wants electric vehicles to be a marketplace success. GM, Ford, Chrysler, electric utilities and government agencies, formed several years ago into the Advanced Battery Consortium, have awarded research and development contracts to five other battery makers pursuing alternative technologies. At the same time, the state Air Resources Board is poised to require automobile companies beginning in 1994 to begin selling a new category of low emitting vehicles that are twice as clean as the cleanest new gasoline cars on the road. Volkswagen will start production of its city car, the Chico, in 1995, and many other manufacturers have similar projects in development. It has attracted attention with the improvements it has made to a proton exchange membrane fuel cell pioneered by General Electric of the US. In that, it differs from other electrical vehicles under development, which are essentially converted delivery vans intended for commercial fleets. It required not only the development of new alloys with the required crash protection properties, but also new production processes. They will be confined mainly to city centres and could be the only kind of car allowed in the most environmentally sensitive areas.

A.6 Ozsoy

Ken Baker, vice-president of GM's research and development centre, insists that 'GM wants electric vehicles to be a marketplace success. RENAULT and Peugeot, the French carmakers, yesterday announced a co-operation accord to help the development of electric cars over the next three years. Electric vehicles are currently uneconomic but California has insisted that carmakers begin offering 'zero emission' vehicles - in other words, electric cars - by 1998 if they are to sell other models in the state. AMERICA'S big three car makers - General Motors, Ford and Chrysler - are to co-operate much more closely in the development of electronic vehicles. The Electric Power Research Institute has worked with both GM and Chrysler to develop electric-powered vans for eventual production. The venture will take place under the auspices of the US Council for Automotive Research (Uscar), an umbrella body which co-ordinates research among the big three, and it will aim to find 'the most effective way to hasten electric vehicle development'. The trio have signed an agreement to investigate co-operation in the design, development, testing and possible manufacturing of electric vehicle components which would ultimately be used in each company's own vehicles.

In the past few weeks Fiat has indicated its intention to produce an electric version of the Cinquecento, its new small car, and Citroen of France has unveiled a prototype electric town car, the Citela.

A.7 Steinberger

So, while a battery vehicle might be practical as a second car, used for short range commuting or shopping, it is a non-starter as an alternative to the family-cum-business car. Yamanouchi's predicts that in 30 years, petrol or diesel-powered cars will account for just 10 per cent of the world's total car output, having been supplanted mainly by hydrogen cars but also by a much smaller proportion of battery powered urban vehicles. Smith refused to estimate how much the car would cost if it went into production, except to say that it would be priced competitively with other cars. Now, nearly a year later, car showrooms clearly are not brimming with alternative-fuel cars. Despite the criticism, electric cars seem poised to represent part of the car market by 1998. Although those gasoline cars are individually dirtier than, say, a car running on a methanol blend, there are more of them. They see electric cars 'as little car for the city only', says Massimello. Unlike other electric vehicles that are conversions of existing cars or vans, GM's version was designed from the ground up as a practical electric car for the consumer market. But, as one Los Angeles car dealer said: 'They can make me put electric cars in my showroom - but they can't make people buy them if they don't want to.'

A.8 QIA-1 and QIA-2¹³

Environmentalists keep on saying the battery-electric is the car of the future. The only question is how soon they will be on the road and what will be in their tanks. Electric cars are the only vehicles to meet such standards so far. How the clean fuels issue will ultimately resolve itself is as murky as the skies over Los Angeles. This does not mean the car-makers are not interested in making electric cars powered by batteries. For at least 30 years, there has been talk of a radically new kind of battery that would make electric cars competitive with petrol or diesel cars. The 200-page study* appears to reinforce the arguments of the US 'big three' car makers, General Motors, Ford and Chrysler, that electric vehicle technology is not sufficiently advanced for viable battery cars to go on sale in California in 1998 in line with state environmental legislation. Unlike other electric vehicles that are conversions of existing cars or vans, GM's version was designed from the ground up as a practical electric car for the consumer market. Were it not for Californian state clean-air legislation requiring 2 per cent of each manufacturer's sales to be of zero-emission vehicles (Zevs) from 1998, it is unlikely that the battery-powered car - currently seen as the only way of achieving zero emissions in urban areas - would be a candidate for volume production this century, certainly in North America.

¹³There was no difference in the extracted summary for this topic