

Automatically Characterizing Salience Using Readers' Feedback

Jean-Yves Delort*

LIRMM,
University of Montpellier 2
delort@lirmm.fr

Abstract

Salience is an important characteristic of information influencing users' cognitive and emotional states. For example, salient parts of a document are those that readers will find moving or provoking. This article analyzes the main characteristics of salience and the different meanings of the concept in information retrieval and linguistics. It also presents a generic approach for identifying linguistically salient segments in a text using readers' textual feedback. The method supports any kind of text and textual feedback. We evaluated the effectiveness of the method with a corpus of blog posts and readers' comments. Our preliminary experiments show that the method has promising results with an fscore of 0.65. The method could also be used on 90% of commented posts which proves that it can be used on a large scale.

Key-words: *Salience, Information Retrieval, Feedback*

1 INTRODUCTION

The most memorable and striking pieces of information in a document are usually those that have caught the most attention. Though relevance is often a factor of salience, it is not a necessary condition of salience. For instance, a reader could remember only one among several relevant piece of information of a text. This article addresses the problem of identifying salient parts in a text and it describes a solution based on readers' textual feedback.

Automatically characterizing salient parts of a text is a difficult problem. First, linguistic salience is not bound to a specific level of linguistic analysis. For instance, a misspelled word, an ambiguous statement or an unbalanced argumentation are typical cases of morphological, semantic and rhetoric salience. Then, the notion of salience is intrinsically contextual. The same statement may be strongly salient in a text and weakly salient in another one. Lastly, linguistic salience is often subjective as it is influenced by past experiences, predispositions and needs.

Salience has received little attention in information retrieval (IR) and knowledge discovery (KD). Most existing works consider relevance to be the main criterion of the importance of information. Relevance usually denotes the usefulness of information to satisfy needs or accomplish a task [9, 28, 36]. Relevance may be assessed by different factors such as [9, 21, 28]: novelty, topicality, informativeness, information quality, usability . . . In the IR field, the terms salience and relevance are often considered as synonyms: *"the terms are used in their literal sense, and often interchangeably, whereas in fact they are distinct notions"* [34].

In contrast, salience has been widely studied in linguistics. One of the main goals of stylistic analysis is to analyze causes and effects of salience of texts [42]. Salience also plays an important role in issues such as generation of referring expressions and anaphora resolution [25, 23]. In this context, the salience of an expression represents its readiness for being related to an object previously introduced in the discourse [15, 14]. Resolution methods have been proposed that compute a degree of salience of an expression. However, these methods do not enable to identify the parts of a text that have caught the most attention.

This article introduces a novel approach for extracting segments (i.e. pre-defined semantic units such as sentence or paragraph) in a document that are salient according to readers' viewpoints. Readers' viewpoints on a document can be derived by analysing textual feedback such as comments or tags. The goal of the proposed method is to locate segments in a text that have been commented in readers' textual feedback. The method starts by identifying pieces of information common to the target and its feedback. They are also weighted according to their size and frequency. Then, the target is segmented and a degree of salience is computed for each segment. Lastly, the segments with the highest values are returned as the most salient ones.

The article is organized as follows. The next section defines the notion of salience with more details and analyses its main characteristics. Section 3 compares the main meanings of salience in linguistics and information retrieval and describes existing methods to evaluate it. Section 4 presents our method for identifying salient segments in a text using readers' textual feedback. We evaluated the effectiveness of the method with a corpus of blog posts and readers' comments. The results are presented in section 5 and they show that the method has promising results with an fscore ranging from 0.55 and 0.65. The method could also be used on 90% of commented posts which proves that it can be used on a large scale.

2 SALIENCE

The salience of an object denotes its capacity to attract or catch an observer's attention. Salience is a general concept that is associated with many everyday terms and expressions (Fig. 1). Salience is an important concept in several

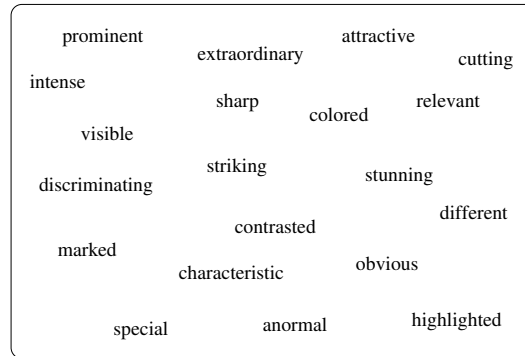


Figure 1: Some terms and expressions related to salience

fields including, computer science, linguistics and psychology. In psychology, attention represents the process that enables organisms to select, among different sources of information, those that will receive cognitive processing. Information is selected according to its salience. Thus, salience denotes a feature of an object (both contextual and subjective) whereas attention is a process. The remaining of this section analyzes the most general characteristics of salience.

Two kinds of salience can be distinguished [24, 34]. Perceptual salience (or physical salience) comes from the perception of the relative prominence of some external features of an object. These features are usually called perceptual or low-level or physical determinants of salience. For example, Aron suggests that accent, intonation and spatial continuity are physical determinants of acoustic salience [1]. Physical determinants of visual salience are, for example, a luminosity contrast, orientation or movement [18]. In contrast, conceptual salience (or cognitive salience) comes from the comparison of the representation of an object with past experiences, context and needs. Conceptual determinants of salience can be *"unexpectedness, unusualness or deviation from the norm"* [34]. Conceptual salience requires semantic processing of the object that may involve interpretation, memory or emotions [2, 24, 31]. Its determinants are usually called "conceptual", "cognitive" or "high-level".

Salience usually involves a combination of conceptual and perceptual determinants that are revealed by concurrent mechanisms [39]. On the one hand, a "bottom-up" mechanism selects stimuli according to salience of their physical characteristics (stimuli-driven). On the other hand, a "top-down" process selects objects according to their semantic salience which depends on the cognitive state of the observer (user-driven).

The salience of an object depends on its context. In his seminal work on similarity measures, Tversky pointed out the importance of context on salience [40]. Tversky's contrast model provides a general framework for the

comparison of similarity measures between two sets of independent properties. A similarity measure between two sets A and B aggregates three values: the degree of inclusion of A in B , the degree of inclusion of B in A and the degree of intersection of A and B . These degrees are computed from a *salience measure* f that reflects the attractiveness of each attribute: " f measures the contribution of any particular (common or distinctive) feature to the similarity between objects". Thus, the salience of the attribute of an object is context-dependent. More specific measures have been proposed to assess the salience of regions in an image [18], sequences in a video [7]. Most of these measures take into account the context.

Salience can be objective or subjective. Typically, conceptual salience is often subjective because it is influenced by past experiences, predispositions and needs. Kamm proposes the following distinction between subjective and objective salience. A thing is subjectively salient if "*it attracts and holds our attention so that we cannot stop thinking about it*" (quoting P. Unger). It is objectively salient if "*it attracts and holds the attention of a normal (or ideal) observer*" [20].

Salience has a social dimension. The social salience of an object represents its perceived distinctive quality and its importance among a group of individuals. Fame, popularity or success are types of social salience which are often determined by citations, links, sells, ... For example, the fact that a researcher has published several articles that have been extensively cited contributes to her fame. In social sciences, media and cultural studies, several kinds of social salience are commonly considered with respect to the size of a society [19, 38]. The social salience of an issue is the importance that it has for a group. Interpersonal salience is the importance that an issue has for someone who discusses it with someone else. Finally, intra-personal salience is the importance of an issue for someone. Group salience refers to "*an individual's awareness of group memberships and respective group differences in an intergroup encounter.*" [16].

3 LINGUISTIC SALIENCE

3.1 Text salience

A text segment may be visually or linguistically salient¹. For instance, visual salience could arise from a color contrast or a change of fontface. The linguistic salience of a segment characterizes its attractiveness from a linguistic viewpoint. Its determinants can be phonetic, morphological, lexical, syntactic, semantic, rhetoric or pragmatic. Figure 3.1 compares a text containing visually salient segments to a text without visual salience.

¹We denote by *segment* a span of text with semantic unity, such as a word, sentence or paragraph.

Query-relevant summarization is an effective technique to extract passages of document that is well-known for its genericity. It is the approach that we have chosen to select commented passages. As explained in Section 2, query-relevant summaries are obtained by scoring sentences with respect to both statistical and linguistic features. To summarize a document, we first split it into sentences. Then, sentences are represented in the vector space model. In this model, a sentence S is represented by a vector of weighted terms: $S = \langle s_i \rangle$, where s_i corresponds to the weight of the term t_i . The weight of a term is given by the number its occurrence within the sentence (term-frequency). Then, a scoring function is used to compare each

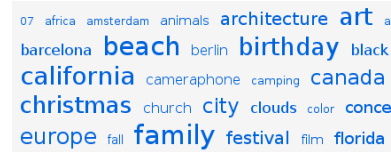


Figure 2: Without visual salience

Figure 3: With visual salience

To appear as linguistically salient to a user, a segment must have been read or at least skimmed. In comparison, visually salient segments draw attention on them almost immediately. As readers' eyes are attracted by visually salient segments, salience in a text often induces non-linear reading. The rest of this section overviews definitions of linguistic salience and existing methods for assessing the salience of a segment.

3.2 Types of linguistic salience

Stylistics is a domain of linguistics that is concerned with the analysis and the description of the goals and effects of distinctive expressions in language. For example, headlines and advertisements are text genres that emphasize important information and are concise and intended to capture attention [12]. The psychological effect produced by a salient textual feature is called *foregrounding* (the term has been borrowed from the visual arts). Foregrounding can be caused either by parallelism or by deviation [26]. Parallelism can be defined as an unexpected regularity in a text. It can be found in constructions such as enumerations, rhymes and alliterations. On the other hand, deviation denotes an unexpected irregularity. A misspelled word, a neologism or a paradox are examples of deviation [27].

From a psycholinguistic perspective, three groups of determinants of salience in a text can be distinguished [24]: physical determinants caused by the form of the text (e.g. syntactic salience caused by the order, and the frequency of occurrences of words), physical determinants caused by the meaning of the text (e.g. salience caused by the topic of the text) and cognitive determinants (e.g. salience caused by emotions).

Salience plays an important role in issues such as generation referring expressions and anaphora resolution [25, 23]. A referring expression is a noun phrase or a pronoun whose function in a text is to "pick out" a place, an event, a person, ... An object that is "picked out" is called a referent. An anaphora is a referring expression whose referent is another referring expression previously mentioned in the discourse. Anaphora resolution consists in identifying in a discourse referents of anaphora of a given type (definite description, pronoun, one-anaphora, ...). In this context, the salience of an expression

represents its readiness for being related to an object previously introduced in the discourse [15, 14].

3.3 Characterization of linguistic salience

Methods for anaphora resolution have been proposed that compute a degree of salience of an expression [25, 22]. They usually assume that the referent is the most salient entity. A degree of salience of an entity is computed using several criteria such as recency, frequency, position in the sentence, ... However, it is not the goal of these methods to identify the parts of a text that capture the most the attention of the reader. Boguraev and Kennedy have adapted one of these methods to identify the most globally salient entities in a text [3]. Their approach tends to identify the entities that have been stressed on by the author which do not necessarily correspond to the most salient entities from the reader's viewpoint. Their approach focuses on physical determinants of salience but it does not take into account cognitive factors.

Numerous works in information retrieval refer to the notion of salience. However, as pointed out by Pattabhiraman and Circone, *"the terms are used in their literal sense, and often interchangeably, whereas in fact they are distinct notions"* [34]. For example, many methods for summarization by extraction are said to be able to identify the most salient segments in a text. However, their real goal is to determine the most informative segments in a text or the most informative with respect to a given topic [3, 13, 11].

Sentiment analysis (or opinion mining) is an important research area that is concerned with the characterization of opinions in texts. One of the main problems is to determine if an opinionated text is positive or negative with respect to a given issue. Methods have been proposed for different types of text (e.g. entity, sentence or document) and different polarities (e.g. binary, discrete or continuous) [32]. Another problem consists in classifying a text as subjective or objective with respect to a given issue. However, even if subjectivity is often a factor of salience, it is not a sufficient criterion for it.

Shipman et al. have studied the probability for an annotated passage to be cited [37]. They have analyzed annotations of law students reading printed case law and writing briefs. Annotations are handwritten and may be textual or markings (e.g. margin bars, circles and underlines). Their work prove that annotated passages are often cited. Recently, Buscher et al. have proposed a method based on an eye-tracking device for determining whether a passage on screen is being read or skimmed [6]. It is possible that combining such a system with an emotion-recognition software could enable not only to identify salient passages but also to characterize the type of reaction they provoke.

Several methods have been proposed to extract passages in a document using readers' feedback [10, 4, 17, 33]. However, these methods are either genre-dependent (e.g. blog posts or product reviews) or feedback-dependent (e.g. annotation or comment).

The next section presents a generic approach for identifying linguistically salient segments in a text which improves the method in [10] and extends it to support any kind of document and textual feedback.

4 PROPOSED APPROACH

In this section, we describe a method for identifying linguistically salient segments in a document using readers' feedback. The expression *textual feedback* denotes a text reflecting readers' reactions to specific or general features of a document (or *target*) that have captured their attention.

Our approach is based on the hypothesis that, *a segment of a text is (subjectively) salient for a reader if he/she reacts to its content*. When a reader is decided to react to information contained in the target, the topics of his/her comment should be close those of that information. Numerous lexical approaches have been proposed to capture and represent the topic of a document. Our approach relies on such techniques to characterize the topics of readers' feedback which are then used to identify segments that are salient for any visual or linguistic reasons (including semantic and pragmatic).

Our method extracts linguistic cues of the salience of segments by identifying all terms and excerpts of text (sequences of terms) that occur both in the feedback and in the target. It involves two steps that are detailed in the sequel of this section:

1. Cue extraction: at this step, all cues are extracted from the feedback,
2. Segment selection: at this step, salient segments are located using previously extracted cues.

4.1 Cue extraction

The first step of the approach consists of extracting salient cues from the feedback. Cue extraction is based on an *N-grams* analysis of the text. Cues are all sub-sequences of terms extracted from reader's feedback with a size ranging from 1 to *N*. Extracted cues are then represented by a weighted vector as in Salton's vector space model [35]. The weight of a cue depends on its size and its number of occurrences.

There are three kinds of cues contained within the target document that our approach can extract from the feedback:

1. Single terms and compound names (e.g. "mobile phone")
2. Named entities (e.g. "Ford Motor Company")
3. Quotes from the original document (e.g. excerpts of a sentence or full sentences).

"I was thinking that"
 "the very last sentence"
 "around the same level"
 "possibility of mobile malware"
 "mobile malware taking down"
 "taking down a phone"

Table 1: Sample of 4-grams generated for a comment

Named entity recognition (NER) is a well-known issue in IR and several methods have been proposed to perform it automatically [25, 29]. Methods have also been proposed to automatically identify compound names [30]. For both problems, N-gram generation has better recall but lower precision than existing methods. However, N-gram generation offers the advantage of being language independent and not requiring machine-learning training. But its main advantage is that it can extract sub-sequences of terms that are readers' quotes on the target. For example, consider the following sentence extracted from a feedback:

I was thinking that this was a decent article until the very last sentence: "The threat level, for now at least, remains at around the same level as the possibility of mobile malware taking down a phone network"².

Table 1 lists a sample of cues of size 4 contained in this sentence. To extract all sub-sequences of words of size 1, ..., N , we use a flexible N-gram analysis tool³.

In the target document, the sentence "The threat level, for now at least, remains at around the same level as the possibility of mobile malware taking down a phone network" contains many N-grams/extracted cues. Note that the salience of this sentence is not due to the choice of words or its syntactic structure, but rather to its meaning. However, the feedback contains lexical cues about it. To put it differently, though the sentence is semantically salient, it could be located by a lexical analysis of the feedback. The degree of salience of this sentence will be determined in the second step.

Once all cues have been extracted, they are assigned a salience weight. The salience function our method utilizes depends on the frequency of the cue in the feedback as well as the cue size, i.e. its number of terms. This approach gives more value to cues involving several terms such as compound names or substrings of quotes from the target. Intuitively, the longer a cue in the feedback, the more likely it is used to refer to an entity of the target. The weight of a cue c is given by $k \times e^{|c|}$, where k is the number of occurrences

²The final part of this sentence is a quote from the target.

³<http://users.cs.dal.ca/~vlado/srcperl/Ngrams/Ngrams.html>

of c in the feedback and $|c|$ is the cue size. The weighted vector of extracted cues represents the *query* Q that will be used in the second step.

Other weighting schemes may be used in order to focus on different aspects of salience. For example, it could be useful to take the number of different readers who have mentioned a cue or the readers' opinions into account.

4.2 Segment selection

The second step of our approach determines the degree of salience of segments that have previously given extracted cues. This problem relates to summarization by extraction, which involves automatically condensing a document. Typically, a summarizer takes a text document as input and it outputs an extract or an abstract. An extract is a selection of the most important sentences of the original document, while an abstract is a summary, at least some items of which do not exist in the original document (e.g. a paraphrase).

Our approach is based on a query-relevant extraction method introduced by Goldstein et al. [13]. Query-relevant methods enable the selection of sentences according to user-defined criteria.

In our approach, segment selection involves three steps. First, the text is fragmented into sentences. Secondly, fragments are represented in a vector space model of cues. The weight of a cue c in a segment v is its number of occurrences v_c . Thirdly, a salience score for each segment is computed and segments with the highest scores are returned to the user as the most salient segments of the text.

In order to compute the salience score of a segment v , given the query Q , we use the *Score* function, which is defined as follows:

$$Score(v) = \sum_{c \in C} v_c \times Q_c$$

where:

- C is the set of cues,
- v_c is the frequency of cue c in segment v ,
- Q_c is the weight of cue c in Q ($Q_c = k \times e^{|c|}$ and k is the cue frequency in the feedback).

4.3 Discussion

Textual feedback may contain only reactions to general features of a document, such as its theme or author. Clearly, such textual feedback cannot be used to identify salient segments. If textual feedback contains reactions to both specific and general features of the target, then the effectiveness of the method will depend on the following soundness criterion: the lower the

lexical intersection of the target with the parts of the feedback that react to general features, the better the effectiveness of the method.

Annotations or comments of a document are not always textual feedback on the document they explicitly link to. For example, blog comments are associated with a post, but they do not necessarily contain reactions to it. Such comments are often textual feedback for another post, or the whole blog (e.g. "I love your blog Bill!"). Therefore, it may be useful to check whether a document is really textual feedback about the document explicitly denoted as its target. An approach to solve this issue using features of the feedback and the context has been proposed in [10].

The proposed method has two important advantages over existing approaches. Firstly, using readers' feedback about a document enables us to identify, with a purely lexical approach, segments that are salient because of visual, phonetic, semantic or pragmatic features. In contrast, existing approaches based on an in-depth analysis of the target content cannot cope with the complexity of processing levels of language higher than syntactic. However, these approaches may not only identify salient segments in a text but also provide reasons for their salience, whereas the proposed approach only focuses on identifying salient segments. Secondly, the method can be used with the feedback of one or several readers which enables to identify socially salient segments. Typically, such segments could contain debated ideas or controversial suggestions.

5 EXPERIMENTATION

In this section, we empirically analyze the effectiveness and scalability of the previous method for identifying salient sentences in a document. The experimentation pursues three objectives. First, it aims at assessing the ability of the method to correctly detect the most salient segments of a text given textual feedback. Secondly, it analyses the correlation between salience and relevance in order to verify the interest of a specific method for characterizing salience in a text. Thirdly, it evaluates the proportion of documents which meet the basic requirements for the method to be used.

This section starts with a corpus overview and then describes the implementation of the approach. Lastly, it presents the three experiments and discusses their main results.

5.1 Corpus features

The experiments were conducted on a corpus of blog posts and their associated comments. In order to build this corpus, we implemented a system that performs the following actions:

1. Query Google to get blog RSS feeds on different topics⁴
2. Download and extract posts from each feed⁵
3. Download all comments associated with each post⁶

The corpus was built with six arbitrarily-chosen queries in order to cover different topics⁷. The corpus contains 11198 posts and 29571 comments. The average number of comments per post is 2.6, but this mean rises to 7.8 when computed from the set of the 3798 posts that have been commented at least once. On average, comments contain 68.3 terms while posts contain 356.7 (for 21.7 sentences on average).

To assess the effectiveness of the proposed method, we will compare sentences selected by our system with gold-standards⁸. Gold-standards are sentences of posts that have been identified by a human expert as salient for at least one commentator. In other words, gold-standards correspond to sentences that have been commented at least once. The selection of posts and identification of gold-standards were carried out with the following process. First, a subset of posts containing at least three sentences and commented by at least three readers is randomly selected from the corpus. Then, the expert reads each post and its comments and annotates sentences of posts corresponding to gold-standards.

Ultimately, the total number of gold-standards is 110, spanning 53 posts. Table 2 summarizes the main corpus features. The corpus can be downloaded at the following URL: <http://www.lirmm.fr/~delort/resources/corpora>

5.2 Implementation

To evaluate the proposed approach, we implemented a system that takes a target document and query document as input and that outputs a ranking of the sentences of the target according to the query. The system performs three steps as part of its execution. The system first extracts N-grams from the content of the target and query. N-grams consisting only of stopwords are filtered out. Secondly, the target is split into sentences. To perform this task, we use a rule-based sentence splitter with an effective disambiguation rate of ". "s larger than 95% [8]. The splitter avoids over-segmentation that would otherwise be triggered by abbreviations (e.g. "*Corp.*") or decimal values (e.g. "44.4"). Thirdly, sentences and query are represented in the

⁴An example of a used query: *inurl:blogspot.com filetype:xml irak war*

⁵An example of an URL of a blog RSS feed: <http://adeshina.blogspot.com/atom.xml>

⁶An example of an URL of a comment feed: <http://fast-weightloss.blogspot.com/feeds/5075260740966721110/comments/default>

⁷Queries: "segolene royale quebec", "straight-edge lifestyle", "chiapas liberalism", "inconvenient truth", "climate crisis", "anna nicole smith"

⁸The presented approach is quite similar to the ROUGE protocol set up by the Document Understanding Conference (DUC).

Number of posts	11198
Average size of a post (number of terms)	356.5
Average number of sentences in a post	21.7
Number of comments	29571
Percentage of commented posts	33.91%
Average number of comments per post	2.6
Average number of comments per commented post	7.8
Average size of a comment (number of terms)	68.3
Total number of gold-standards	110
Total number of annotated posts	53

Table 2: Corpus features

vector-space model. As previously explained, the weight of a cue in the query depends on its frequency and size whereas the weight of a cue in a sentence only depends on its frequency. Finally, the *Score* function defined in 4.2 is used to compare each sentence with the query. When the query is the concatenation of the comments of a post, the system will output a list of sentences of the target ranked with respect to function *Score* which gives their degree of salience.

5.3 Results and Discussion

5.3.1 Effectiveness

We compare the effectiveness of the proposed approach (in the following called "comment method") using different N-gram values: 1, 2, 3 and a high value, 8, which prioritize sentences in targets that contain longer cues. Effectiveness is assessed with the following three criteria that are standard for this kind of evaluation: precision, recall and fscore [41]. In our problem, they are defined as:

$$\text{Precision} = \frac{\text{total number of gold-standards selected by the system}}{\text{total number of sentences selected by the system}}$$

$$\text{Recall} = \frac{\text{total number of gold-standards selected by the system}}{\text{total number of gold-standards}}$$

$$\text{Fscore} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

A common problem in the evaluation of a ranking method is the way to deal with objects that have the same values. For example, this situation

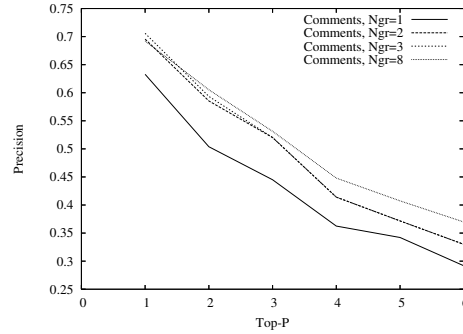


Figure 4: Precision

occurs if a user wishes to know only the most salient sentence and if two sentences have the highest value. The system could return both sentences or it could randomly take one of them. We consider here that if the method has not returned the maximum number of required sentences, and the set of sentences with the highest value is not unique, then the entire set should be selected. This approach lowers the precision because more sentences will be selected and, at the same time, it increases the recall. In the experimentation context, it is still a more suitable approach than randomly choosing a sentence, as this would introduce more noise. P denotes the number of sentences that have to be selected.

Figures 4 and 5 present the precision and recall for the comment method (results are also detailed in Annex).

The precisions of the generic method with different N values are almost the same. They remain nearly constant for up to two selected sentences, then they decline due to the gap between P and the average of gold-standards. The Comment method has twofold better precision values than the Generic method.

First, we can see that using N -grams bring a 15% improvement compared to a simple bag-of-words approach (when $N = 1$). Precision is also good when the method returns only one result ($P = 1$). However, as few posts contain more than one gold-standard, precision decreases at P increment. Precision lowers also because of chosen sentence overselection policy.

Recall has a steeply increasing slope for the Comment when P ranges from 1 to 3, then its slope softens but stays positive. When $P = 3$, the recall is 0.82 for N -grams of size 2 or 3. Rappel slowly increases of 0.1 between $P = 3$ and $P = 6$. Figure 6 presents a comparison of the overall effectiveness of the comment method for different P and N values.

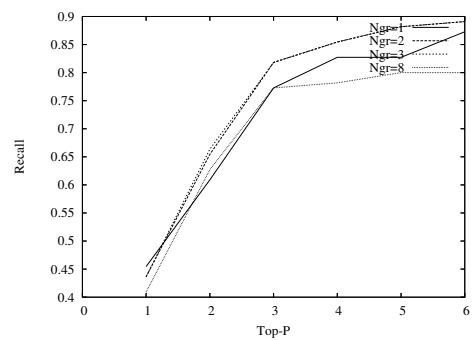


Figure 5: Recall

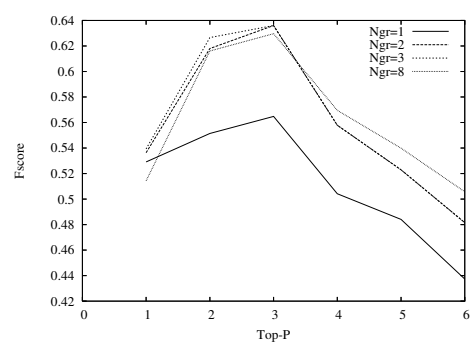


Figure 6: F-score

Several remarks may be made about posts containing gold-standards. First, some of them contain questions or requests which probably invite the readers to respond. Secondly, they often present lists of preferences, alternatives or different points of view. In that case, readers tend to stand for a specific position or idea. Lastly, commented documents often belong to the same blogs which possess an audience made up of regular readers. Regular readers comment more frequently than occasional readers because they are *a priori* more interested in the blog.

5.3.2 Correlation relevance - salience

It can be interesting to check if the most salient sentences are also the most relevant ones. Indeed, extraction techniques for summarization have been proved efficient to identify the most relevant segments in a document. If the proportion of sentences which are both salient and relevant was important, then it would not be necessary to resort to a specific method to extract salient sentences.

In order to evaluate the correlation relevance-salience, we compare the most relevant sentences of the document with the gold-standards. The evaluation is indirect because the most relevant sentences are not identified by human experts, but by an extraction technique for summarization adapted to produce generic summaries.

In order to extract the most relevant sentences, we resort to the system presented in 5.2 providing it as input query, the target document. Several N-grams sizes will be also tested. Precision and recall for this method are presented in figure 7. We observe that results are in average 50% lower than with the comment method. Assuming that Goldstein's method perfectly identifies the most relevant sentences, we can deduce that less than 30% of sentences are both the most salient and the most relevant. In other words, there is not a strong correlation between salience and relevance. This confirms the specificity of the problem of characterizing salient segments in a document.

5.3.3 Usability

We now analyze the usability of the proposed approach. A sufficient condition for the usability of the method on a document is that cues of target segments occur in the feedback. In our setting, we need to determine the proportion of posts that contain some cues of their comments. We focus on the size of the longest cue, in other words, on the size of the longest sequence of terms shared by the post and the comments. Figure 8 presents the results for the 3798 posts of the corpus which have been commented at least once.

Only 10.4% of the posts have no intersection at all with their comments. In most posts (72.3%), the size of the longest cue size is 1 or 2 (e.g. "Hillary Clinton"). Thus, 17.2% of posts contain a long cue (with a size ranging from

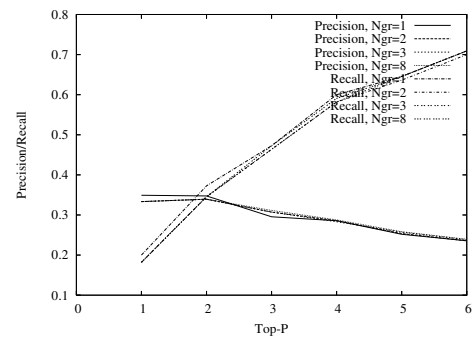


Figure 7: Precision and recall for a generic summary

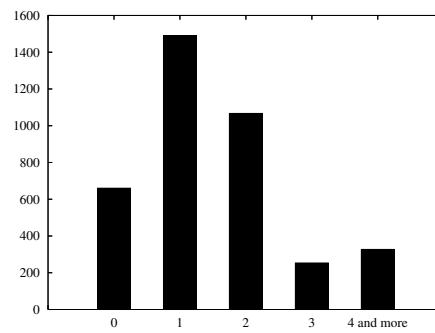


Figure 8: Distribution of the longest cue size in targets

3 to 10) in their feedback. These results highlight the usability of the method, which can be applied to more than 90% of considered documents. In the case where the lexical intersection between the target and the feedback is empty or almost empty, there vector representations can be enhanced in order to broaden their lexical fields and make them overlap. Different techniques may be useful for broadening a lexical field such as using a thesaurus [43] or taking into account pseudo-feedback [5].

6 SUMMARY AND FUTURE RESEARCH

After analyzing the main difficulties of content-based techniques to automatically identify salient segments in a document, we have presented a new context-based method. Our approach analyzes the readers' textual feedback about a document to extract what they find salient in it. A first limitation of this approach is that it cannot be used on documents without feedback. Another limitation is that it does not characterize the type of salience associated to a segment. The experiments on a blog post corpus have shown promising results.

The method could be useful with other types of documents. For example, it could be interesting to use the method with newspaper articles or wiki pages. Typically, some passages of the Wikipedia page about "Genetically modified organisms" are actively debated in the associated discussion page. The proposed approach should be able to identify them. The method could also be used to locate and correct ambiguous or unclear passages in technical documentations using readers' feedback. Lastly, another type of corpus are contracts. Contracts are interesting documents because their authors tend to deliberately dissimulate relevant information such as unfair terms. The proposed method could be able to locate such terms by processing clients' feedback.

REFERENCES

- [1] B. Arons. A review of the cocktail party effect. *Journal of the American Voice I/O Society*, 12:35–50, July 1992.
- [2] Erik Blaser, George Sperling et Zhong L. Lu. Measuring the amplification of attention. *Proceedings of the National Academy of Sciences of the United States of America*, 96(20):11681–11686, September 1999.
- [3] B. Boguraev et C. Kennedy. Salience-based content characterisation of text documents. In *ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, 1997.
- [4] Oisín Boydell et Barry Smyth. From social bookmarking to social summarization: an experiment in community-based summary generation.

In *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*, pages 42–51, New York, NY, USA, 2007. ACM Press.

- [5] Chris Buckley, Gerard Salton, James Allan et Amit Singhal. Automatic query expansion using SMART: TREC 3. In *Proceedings of Text REtrieval Conference*, pages 69–80, 1994.
- [6] Georg Buscher, Andreas Dengel, Ludger van Elst et Florian Mittag. Generating and using gaze-based document annotations. In *CHI '08: CHI '08 extended abstracts on Human factors in computing systems*, pages 3045–3050, New York, NY, USA, 2008. ACM.
- [7] Guillaume Cabanac, Max Chevalier, Claude Chrisment et Christine J. 0002. A social validation of collaborative annotations on digital documents. In Jean F. Boujut, éditeur, *IWAC*, pages 31–40. CNRS - Programme société de l'information, 2005.
- [8] Paul Clough. A perl program for sentence splitting using rules. Technical report, University of Sheffield, 2001.
- [9] E. Cosijn et P. Ingwerson. Dimensions of relevance. *Information Processing & Management*, 36:533–550, 2000.
- [10] Jean Y. Delort. Identifying commented passages of documents using implicit hyperlinks. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 89–98, New York, NY, USA, 2006. ACM Press.
- [11] G. Erkan et Dragomir R. Radev. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22, 2004.
- [12] Petr Forst. Stylistic analysis and interpretation of linguistic patterns in advertisements. Master's thesis, Univerzita Pardubice, 2008.
- [13] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal et Jaime Carbonell. Summarizing text documents: sentence selection and evaluation metrics. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 121–128, New York, NY, USA, 1999. ACM Press.
- [14] Barbara J. Grosz, Scott Weinstein et Aravind K. Joshi. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.
- [15] Eva Hajičová et Petr Sgall. Topic-focus and salience. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 276–281, Morristown, NJ, USA, 2001. Association for Computational Linguistics.
- [16] Jake Harwood, Priya Raman et Miles Hewstone. The family and communication dynamics of group salience. *Journal of Family Communication*, 6(3):181–200, 2006.

- [17] Meishan Hu, Aixin Sun et Ee P. Lim. Comments-oriented blog summarization by sentence extraction. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 901–904, New York, NY, USA, 2007. ACM.
- [18] L. Itti, C. Koch et E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.
- [19] Owen D. Jones. Characterising the social salience of electronically mediated communication. In *CHI '94: Conference companion on Human factors in computing systems*, pages 93–94, New York, NY, USA, 1994. ACM Press.
- [20] F. M. Kamm. Does distance matter morally to the duty to rescue. *Law and Philosophy*, pages 655–681, November 2000.
- [21] J. Kekäläinen et K. Jäärvelin. Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In H. Bruce et R. Fidel, éditeurs, *Proceedings of CoLIS 4 conference*, pages 253–270, Greenwood Village, 2002.
- [22] Christopher Kennedy et Branimir Boguraev. Anaphora for everyone: pronominal anaphora resolution without a parser. In *Proceedings of the 16th conference on Computational linguistics*, pages 113–118, Morristown, NJ, USA, 1996. Association for Computational Linguistics.
- [23] E. Krahmer et M. Theune. Efficient generation of descriptions in context. In *Proceedings of the ESSLLI workshop on the generation of nominals*, 1999.
- [24] F. Landragin. Saillance physique et saillance cognitive. *Cognition, Représentation, Langage*, 2(2), 2004.
- [25] Shalom Lappin et Herbert J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, December 1994.
- [26] Dan McIntyre. Using foregrounding theory as a teaching methodology in a stylistics course. *Style*, 37(1):1–13, 2003.
- [27] Masanori Miyata. Types of linguistic deviation in oliver twist. *Bulletin of Shikoku Women's University*, 1(1):1–18, 1981.
- [28] Stefano Mizzaro. How many relevances in information retrieval? *Interacting with Computers*, 10(3):303–320, 1998.
- [29] D. Mollá, R. Schwitter, F. Rinaldi, J. Dowdall et M. Hess. Anaphora resolution in extrans. In *Proceedings of the International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization*, pages 67–74, 2003.
- [30] Goran Nenadic et Irena Spasic. Recognition and acquisition of compound names from corpora. In *NLP '00: Proceedings of the Second*

International Conference on Natural Language Processing, pages 38–48, London, UK, 2000. Springer-Verlag.

- [31] Paula M. Niedenthal et S. Kitayama, éditeurs. *The heart's eye: Emotional influences in perception and attention*. New York: Academic Press, 1994.
- [32] Bo Pang et Lillian Lee. Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, 2(1-2):1–135, 2008.
- [33] Jaehui Park, Tomohiro Fukuhara, Ikki Ohmukai, Hideaki Takeda et Sang G. Lee. Web content summarization using social bookmarks: a new approach for social summarization. In *WIDM '08: Proceeding of the 10th ACM workshop on Web information and data management*, pages 103–110, New York, NY, USA, 2008. ACM.
- [34] T. Patabhraman et Nick Cercone. Selection: Salience, relevance and the coupling between domain-level tasks and text planning. In *Proceedings of the Fifth International Workshop on Natural Language Generation*, pages 79–86, 1990.
- [35] G. Salton, A. Wong et C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, 1975.
- [36] Tefko Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science, part ii: nature and manifestations of relevance. *Journal of the American Society of Information Sciences and Technologies*, 58(13):1915–1933, 2007.
- [37] F. M. Shipman, M. N. Price, C. C. Marshall et G. Golovchinsky. Identifying useful passages in documents based on annotation patterns. In *Proceedings of the European Conference on Digital Libraries*, pages 101–112, 2003.
- [38] G. Stephenson. Intergroup bargaining and negotiation. In J. Turner et H. Giles, éditeurs, *Intergroup behavior*, pages 168–198. Chicago: University of Chicago Press, 1981.
- [39] Anne Treisman. Features and objects in visual processing. *Scientific American*, 255(5):114–125, 1986.
- [40] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.
- [41] C. J. Van Rijsbergen. *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow, 1979.
- [42] Peter Verdonk. *Stylistics*. Oxford University Press, 2002.
- [43] Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

ANNEX

Method	1	2	3	4	5	6
Comment Ngr=1	0.529	0.551	0.565	0.504	0.484	0.437
Comment Ngr=2	0.536	0.618	0.636	0.558	0.523	0.482
Comment Ngr=3	0.539	0.627	0.636	0.558	0.523	0.482
Comment Ngr=8	0.514	0.616	0.630	0.570	0.540	0.506
Generic Ngr=1	0.254	0.360	0.364	0.387	0.362	0.354
Generic Ngr=2	0.235	0.342	0.370	0.383	0.363	0.352
Generic Ngr=3	0.235	0.342	0.370	0.381	0.368	0.357
Generic Ngr=8	0.235	0.342	0.375	0.387	0.369	0.358

Table 3: F-score

Method	1	2	3	4	5	6
Comment Ngr=1	0.455	0.609	0.773	0.827	0.827	0.873
Comment Ngr=2	0.436	0.655	0.818	0.855	0.882	0.891
Comment Ngr=3	0.436	0.664	0.818	0.855	0.882	0.891
Comment Ngr=8	0.409	0.627	0.773	0.782	0.800	0.800
Generic Ngr=1	0.200	0.373	0.473	0.600	0.645	0.709
Generic Ngr=2	0.182	0.345	0.464	0.582	0.636	0.700
Generic Ngr=3	0.182	0.345	0.464	0.582	0.645	0.709
Generic Ngr=8	0.182	0.345	0.473	0.591	0.645	0.709

Table 4: Recall

Method	1	2	3	4	5	6
Comment Ngr=1	0.633	0.504	0.445	0.363	0.342	0.292
Comment Ngr=2	0.696	0.585	0.520	0.414	0.372	0.330
Comment Ngr=3	0.706	0.593	0.520	0.414	0.372	0.330
Comment Ngr=8	0.692	0.605	0.531	0.448	0.407	0.370
Generic Ngr=1	0.349	0.347	0.295	0.286	0.252	0.236
Generic Ngr=2	0.333	0.339	0.307	0.286	0.254	0.235
Generic Ngr=3	0.333	0.339	0.307	0.283	0.257	0.239
Generic Ngr=8	0.333	0.339	0.311	0.288	0.258	0.239

Table 5: Precision