On the Subjectivity of Human Authored Short Summaries

BalaKrishna Kolluru   Yoshihiko Gotoh
University of Sheffield, Department of Computer Science
Sheffield S1 4DP, United Kingdom
{b.kolluru, y.gotoh}@dcs.shef.ac.uk

Abstract

We address the issue of human subjectivity when authoring summaries, aiming at a simple, robust evaluation of machine generated summaries. Applying a cross comprehension test on human authored short summaries from broadcast news, the level of subjectivity is gauged among four authors. The instruction set is simple, thus there is enough room for subjectivity. However the approach is robust because the test does not use the absolute score, relying instead on relative comparison, effectively alleviating the subjectivity. Finally we illustrate the application of the above scheme when evaluating the informativeness of machine generated summaries.

1 Introduction

Subjectivity plays an important role when removing the unwanted or redundant information for summarising a document. Human beings tend to disagree on what should be a 'one good summary' (Mani, 2001). This is probably because every individual, whilst arriving at a summary, looks at things from a different perspective. Guided by various factors such as educational background, profession, personal interests and experience, an individual decides whether a certain aspect is worth being included in a summary. What might seem relevant to one person could be deemed redundant by another when reading the same story, thus accounting for more than one 'correct' summary. The issue of subjectivity gains prominence as the compression ratio increases, *i.e.*, the shorter the summary, the larger the number of 'correct' summaries (Lin and Hovy, 2003b). This is due to the fact that assimilation of seemingly important contents takes priority while discarding the redundant information. This is a highly subjective aspect.

Although the subjectivity reflects individual's thoughts, there will also be some information commonly observed in different summaries of the same story. Stated otherwise, words in a summary may vary, phrases may vary, and often the grammatical structure may not be the same, but a certain degree of information may be common across summaries. To what degree is information uniform across different summaries? How much subjectivity is there? How do we account for similar information stated using different words, expressions, or grammatical structure when comparing summaries? How does this help when gauging the informativeness? Does the subjectivity cause any adverse effects when evaluating summaries? It is these questions that we aim to address in this paper.

Let us assume that the atomic facts of a summary account for its relevance. Then, a simple question that elicits any one of these atomic facts represents a benchmark for assessing its informativeness. We wish to evaluate the quality of a summary in terms of atomic facts commonly observed in-, or subjectively discarded from, assorted human authored short summaries. In our quest to quantify the subjectivity, we devise a cross comprehension test along the lines of (Hirschmann et al., 1999) for extracting atomic contents. The comprehension test is modelled on a question-answer style framework. 'Crossing' the model turns out to be an effective scheme for measuring the divergence among multiple summaries. Questions are prepared by the subject who wrote the original summary (Section 3). Their answers

should be derived by reading the summary alone. Summary-questionnaire pairs are then swapped in such a way that any summary is paired with questions written by other subjects (Section 4). The number of questions that cannot be answered by reading the summary accounts for the subjectiveness of the author (Section 5). Finally, we address how the cross comprehension test can be used for evaluating machine generated summaries (Section 6).

2   Related Works

There have been a number of studies concerned with collating and analysing of human authored summaries, with the aim of producing and evaluating machine generated summaries. A phrase weighting process called the 'pyramid method' was described in (Nenkova and Passonneau, 2004). They exploited the frequency of the same (similar) information that was in multiple summaries of the same story. It was referred to as a *summarisation content unit* (SCU). Increasing stability of pyramid scores was observed as the pyramid grew larger. The authors concluded, however, that the initial creation of the pyramid was a tedious task because a large number of SCUs had to be hand annotated.

In (Van Halteren and Teufel, 2003), the co-occurrence of atomic information elements, called *factoids*, was examined whilst analysing 50 different summaries of two stories. A candidate summary was compared with the reference using factoids in order to measure the informativeness. The authors observed that from a wide selection of factoids only a small number were included in all summaries. From a pool of factoids, approximately 30% were taken to build a consensus summary that could be used as a 'gold standard'.

Summary evaluation has been recognised as a sensitive, non-trivial task. In (Radev and Tam, 2003) the *relative utility* was calculated based on a significance ranking assigned to each sentence. A word network based summary evaluation scheme was proposed in (Hori et al., 2003), where the accuracy was weighted by the posterior probability of the manual summaries in the network. Significantly, they surmised the independence of their criterion from the variations in hand summaries.

A regression analysis was performed in (Hiro-

hata et al., 2005) and concluded that objective evaluations were more effective than subjective approaches. Although their experiments were concerned with presentation speech, the results do have a universal appeal.

Another notable development in the field is the n-gram co-occurrence matching technique as proposed in (Lin and Hovy, 2003a). Their tool, ROUGE, compares the number of n-gram matches between a reference and a machine generated summary. Recently, ROUGE was piloted for evaluation of summaries from newspaper/newswire articles (Over and Yen, 2004). ROUGE simulated the manual evaluation well for that task, although it is still unclear how closely it well to other tasks.

To some extent, the work described in this paper is close to that of (Nenkova and Passonneau, 2004) and (Van Halteren and Teufel, 2003). We analyse human authored summaries associating human subjectivity with their unique interpretation of stories. We consider their effect when evaluating machine generated summaries.

3   Production of Human Authored Short Summaries

Our aim is to investigate an effective, robust approach to summary evaluation. In this paper, we identify and quantify the aspect of human subjectivity while authoring short summaries. To this end, four subjects produced a short summary (approximately 100 characters, or 15 words) for broadcast news stories given a simple instruction set. This summary is referred to as a 'one line' summary because it corresponds approximately to the average sentence length for this data set.

3.1   Author Profiles

Four summary authors are briefly profiled below:

Subject A. A linguist by profession, a polyglot out of interest, and an author by hobby. This subject is fluent in English, Spanish and French; English being the first language. The subject is trained to write summaries and translations.

Subject B. A manager by qualification and a polyglot by necessity; English is a second language. This subject was trained in making presentations and documentation. We hoped to benefit from the synergy

of both fields for summary production.

Subject C. A physicist by qualification and currently working towards a PhD in speech recognition. English is the first language. In addition, this subject has an interest in theatre and drama, thus is exposed to literature and related fields.

Subject D. Working on research in multiparty meetings as a post doctoral fellow. English is the first language for this subject. Experience of meeting summarisation.

All subjects are educated to at least graduate level, and have are fluent in English. It was expected that they could produce summaries of good quality without detailed instruction or further training. A simple instruction set (discussed later) was given, leaving wide room for interpretation about what might be included in the summary. Hence subjectivity was promoted.

3.2 Data

The human subjects worked on a small subset of American broadcast news stories from the TDT-2 corpus (Cieri et al., 1999). They were used for NIST TDT evaluations and the TREC-8 and TREC-9 spoken document retrieval evaluations. Each program in the corpus contained 7 to 8 news stories on average, spanning 30 minutes as broadcast which might be reduced to 22 minutes once advertisement breaks were removed. A set of 51 hand transcriptions were manually selected from the corpus. The average length was 487 words in 25 sentences per transcription.

3.3 Instructions

Summary production. A simple instruction was given to the subjects in order to arrive at a summary:

- Each summary should contain about 100 characters, possibly in the subject's own words.

As the news stories ranged from 16 to 84 sentences, subjects would have to prioritise information that could be included in their 'one line' summary. The instruction implicitly encouraged the subjects to put as much important information as possible into a summary, while maintaining a good level of fluency. It was also a flexible instruction so that subjects were able to use their own expressions when necessary. After completion of the task, they commented that

this instruction made them experiment with different words to shorten or expand the information they wanted to include. For example, how could an earthquake disaster be expressed in different ways:

> 8000+ feared dead? . . . or
> thousands of people killed? . . . or
> a lot of people are believed to be dead?

Another feature of this instruction was the amount of generalisation that a subject was likely to use. For example, a subject could say

> US Senate to decide on tobacco bill

but given the length constraints, it could be like

> Senate to vote on bill, hiking tobacco price

while adding extra information, but omitting specific details.

Questionnaire production. When producing summaries, subjects were aware that they also had to prepare questions with the following instructions:

- A questionnaire may consists of 2–4 questions;

- An answer must be found in the particular summary, without reading the entire story;

- Yes / no questions should not be used;

- The summary may roughly be reconstructed from the question-answer set.

Each fact might be questioned in such a way that the particular summary could be recovered. Ideally we would expect each question to elicit a precise information point chosen for the summary — *e.g.*, who did it, when did it happen, what was the cause? The question-answer set enabled us to gauge the most relevant information as decided by the subjects, so that their subjectiveness became apparent.

3.4 Full Sample

A 'one line' summary-questionnaire pair was produced for 51 broadcast news stories by each of the four subjects. The statistics in Table 1 show the average number of words and characters for each summary. It is observed that Subjects **A** (6.1 characters / word) and **C** (5.8) tended to use longer words than **B**

Subject	#words	#characters	#questions
A	16	113	3.7
B	17	99	3.5
C	12	81	2.4
D	21	131	3.0

Table 1: This table shows the average number of words and characters for each summary, and the average number of questions per summary.

(4.9) and **D** (5.3). The table also shows how the average number of questions varies between subjects.

Table 2 shows a full sample. The complete news story is found in the Appendix. The difference between the four summaries can be clearly observed. One noticeable aspect is the amount of abstraction preferred by various subjects. Both Subjects **A** and **D** fully utilised words from the news story and made a small amount of abstraction. In particular, Subject **A** chose to pick out a person ('*Fisher*') who conducted the study, while **D** opted for specifics of the study ('*dopamine*' — a responsible chemical). On the other hand, Subjects **B** and **C** have rendered their interpretation of the story in their own expressions. They have produced a highly abstracted summary reflecting the sense of the story while ignoring the specifics — nevertheless they were very different from each other. All four summaries happen to be of good quality, however it is the sheer divergence in the words, the expressions and subjective interpretation that is striking.

Word usage among the subjects is also interesting — *e.g.*, '*visual images*' as against '*physical traits*'; similarly '*inner feelings*' as against '*chemistry*'. Such expressions and idioms are open for interpretation, making it difficult to quantify the informativeness of any summary.

There also exist many factual news stories among the 51 test stories. It is left for a future study to compare between factual and non-factual news, in particular about the amount of abstraction.

4 Cross Comprehension Test

Each question can extract a relevant answer from the particular summary by the same author. If a question set were applied to a different summary, some answers may be discernible whereas others may not. The cross comprehension test achieves this by swap-

---

Subject A
Summary:
Fisher's study claims we seek partners using unconscious love maps; women prefer status, men go for physical traits.
Questions:
1. Who is the author of this study?
2. What claim does the researcher make concerning our method for seeking a sexual partner?
3. What do women look for in men?
4. What do men go for?

Subject B
Summary:
Internal feelings of love between men and women are unique; external features depend on culture.
Questions:
1. What are unique?
2. What is this topic about?
3. What differs between men and women?
4. Why does it differ?

Subject C
Summary:
Culture and chemistry both play a role in the science of romance.
Questions:
1. What is being discussed?
2. What are the factors affecting the particular event?

Subject D
Summary:
Men are turned on by visual images and women are more focused on someone's character traits, based on dopamine.
Questions:
1. What do women look for in men?
2. What do men look for in women?
3. What is the chemical that controls attraction?

Table 2: Summary-questionnaire pairs produced from broadcast news stories by four subjects.

ping a summary-questionnaire pair, *i.e.*, each summary was paired with questions produced by different authors. Figure 1 illustrates the way it works.

A single judge examines whether each question can be answered by reading a swapped summary. The judge is a person different from the four summary authors. Further, if the answer is found, it may be relevant, partially relevant, or totally irrelevant to the one expected by the author. Thus, the decision is made from the following four options:

relevant: a relevant answer is found — the answer is deemed to be relevant if it conveys the same meaning as expected by the author even if a different expression is used;

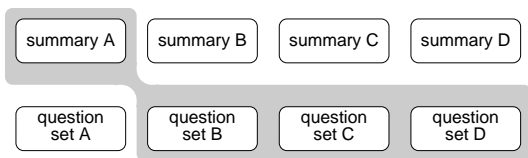*partially relevant*: an answer is partially relevant;

Figure 1: The cross comprehension test swaps summary-questionnaire pairs between subjects. For example, a summary by Subject **A** may be questioned by those set by Subjects **B**, **C**, and **D**.

irrelevant: an answer is found, but is totally different from that expected by the author.

not found: no answer is found.

Sample (re-visited). Table 3 shows the summary and questions crossed from the sample in Table 2. For example, when the 'one line' summary authored by Subject **A** is matched with Subject **B**'s questions, corresponding answers may be

1. ?;
2. seeking partners;
3. women prefer status, men go for physical traits;
4. unconscious love maps.

We may thus conclude answers are '*not found*', '*relevant*', '*irrelevant*', and '*partially relevant*' because, from Table 2, actual answers sought by **B** were

1. internal feelings;
2. love between men and women;
3. external features;
4. cultural reason.

Compensating ill-framed questions. We are aware that not all 'one line' summaries were well written. For example, it may be difficult to reach the expected answer ('external features') for Question 3 by Subject **B** ('What differs between men and women?') by reading the summary from the same subject. Moreover, subjects occasionally set a question that could not be answered properly by reading the particular summary alone. By crossing the summary-questionnaire pair, ill-framed questions are effectively compensated, because they are equally posed to all candidate summaries.

Judgement difficulty. One potential problem in this scheme is the difficulty a judge may face when choosing from the four options. A judge's decision can also be affected by subjectivity. Our assumptions are that (1) because there are only four options, there is less room for the subjectivity in comparison

Summary by Subject A:
Fisher's study claims we seek partners using unconscious love maps; women prefer status, men go for physical traits.

Questions by Subject B:
1. What are unique?   (N)
2. What is this topic about?   (R)
3. What differs between men and women?   (I)
4. Why does it differ?   (P)

Questions by Subject C:
1. What is being discussed?   (R)
2. What are the factors affecting the particular event?   (R)

Questions by subject D:
1. What do men look for in women?   (R)
2. What do women look for in men?   (R)
3. What is the chemical that controls attraction?   (N)

Table 3: What if the summary by Subject **A** is questioned by Subjects **B**, **C**, or **D**? (R), (P), (I), and (N) after each question indicate the answer is *relevant*, *partially relevant*, *irrelevant*, and *not found.*

to the summary writing task, and that (2) a decision between '*relevant*' and '*partially relevant*' and one between '*irrelevant*' and '*not found*' are both not very important because the former two are roughly associated with commonly shared information and the latter two correspond to the subjective part. Although the following section shows results by a single judge, we are currently conducting the same experiments using multiple judges in order to quantify our assumptions.

5 Evaluation Results

Each of the four 'one line' summaries from the 51 broadcast news stories were evaluated using three sets of 'crossed' questions.

5.1 Summary Relevance

Figure 2(a) shows, when paired with questions by other subjects, how many answers could be found in a candidate summary. The figure indicates that summaries authored by the different subjects contained '*relevant*' information for less than half (47% overall average for four subjects) of questions. The number goes up slightly (61%) if '*partially relevant*' answers are included. The number of answers that were '*not found*' indicates the level of subjectivity for this 'summary writing' exercise; more than one third (35%) of information that one subject thought
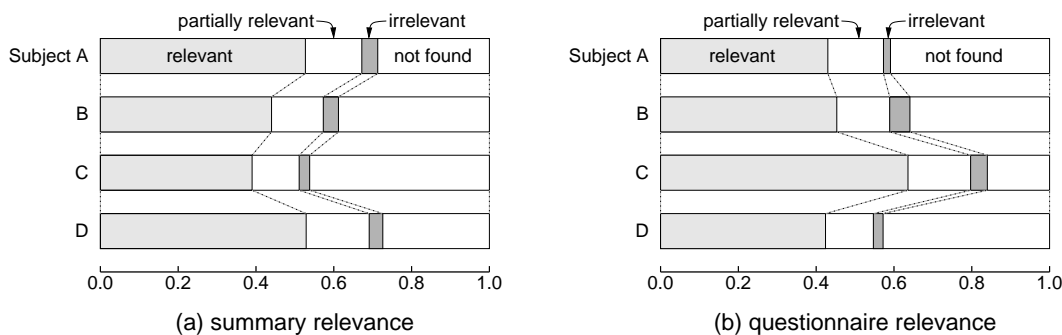
Figure 2: Summary relevance was measured when evaluated against questions by other subjects, while questionnaire relevance was calculated when evaluated against summaries by other subjects.

was the most important was discarded by the others. We surmise that '*irrelevant*' answers were also caused by the subjectivity; occasionally authors arrived at contradictory summaries of the same story due to its ambiguous nature. In such cases, questions were produced from that author's subjective view, and they certainly affected the relevance of a summary by the other subject.

Another notable outcome of this experiment is that the number of answers found '*relevant*', '*partially relevant*' or '*irrelevant*' was 71%, 61%, 54% and 73% for Subjects **A**, **B**, **C**, and **D**, respectively. This seems roughly proportional to the average length of summaries by each subject (113, 99, 81, and 131 characters, respectively). The longer the summary, the more information one can write in the summary. It is thus hypothesised that only the summary length matters for finding the '*relevant*' information in summaries. Looking at this outcome from a different perspective, there is no evidence that one author was more subjective than the others.

5.2 Questionnaire Relevance

Figure 2(b) shows, when paired with summaries by other subjects, how many candidate questions could be answered. It is based on the same evaluation as 2(a), but observed from the different angle. Approximately the same number (55–59%) of '*relevant*', and '*partially relevant*' answers were found for Subjects **A**, **B**, and **D**. However, it was much higher (80%) for Subject **C**. The reason seems to be that this subject frequently set questions that might accept a wide range of answers, while other subjects tended to frame questions that required more

specific information in the summary; *e.g.*, Subject **C**'s '*what is being discussed?*' was a general question that was more likely to have some answer than Subject **B**'s question '*what differs between men and women?*'.

5.3 Discussion

The overall number of '*relevant*' and '*partially relevant*' answers found by the cross comprehension test was just over 61% for four subjects. This accounts for the amount of information that was agreed by all the subjects as important. For more than one third of summary contents, subjects had different opinions about whether they should be in their 'one line' summaries, resulting in categories such as '*irrelevant*' or '*not found*'. Occasionally these categories resulted from ill-framed questions, but such questions were infrequent. For most of the cases, they were caused by the subjectivity of a different individual.

We noted earlier that only the summary length matters and there is no evidence that one author was more subjective than the others. It is probably because, given a clear instruction about the summary length (*i.e.*, roughly 100 characters for this task), there is an upper bound for the amount of information that anyone can fit into the summary, while maintaining fluency. When the summary is short, one has to make a serious decision about which important information should go into a summary, and the decision often reflects one's subjective thoughts. Our argument is that, assuming the subject's effort, the amount of subjectivity was controlled by the summary length constraints rather than an individual's nature.
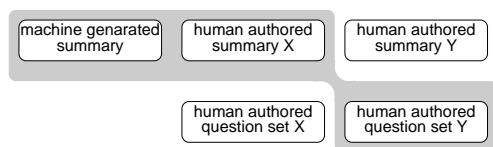
14

Figure 3: Evaluation of machine generated summaries by the cross comprehension test.

The diversity of summaries caused by individual subjectivity may be alleviated by carefully drafting an instruction set. However it probably results in a large list of instructions, and the drafting process certainly will not be straightforward. Further, it is not likely that we can ever completely remove the subjectivity from human work. Indeed, if subjectivity disappeared from human authored summary by well crafted instructions, it would be more like turning human activity into a mechanical process, rather than a machine to simulate human work.

A non-trivial problem of the approach may be the amount of human effort needed for evaluation. Production of summary-questionnaire pairs may not be difficult, as it is based on a simple instruction set and even accepts ill-framed questions, but it still requires human time. On the other hand, a judge's role is the most critical — it is labour intensive, and the effect of potentially subjective judgement needs to be studied.

Although certainly not flawless, the cross comprehension test has its own advantage. A simple instruction set is effective; it encourages authors to make their best effort to put as much information into a short summary. Most importantly, the test is robust; it sometimes causes ill-framed questions, but they can be compensated by relative comparison achieved by crossing summary-questionnaire pairs.

6 Evaluation of Machine Generated Summaries

The objective of this evaluation is to measure the information content of machine generated summaries using a human authored summary as a yardstick. Although very subjective for many cases, a human summary can still be a reference if we do not treat them as a 'gold standard'.

The cross comprehension test of machine generated and human authored summaries is illustrated in

---

> **Machine generated summary:**
> senate to vote to approve the expansion of north atlantic treaty organisation to bigger nato means us obligations
>
> **Summary by subject B:**
> US Senate to decide on NATO expansion; US assesses bigger NATO more arms deal but poor ties with Russia.
>
> **Questions by subject D:**
> 1. What is happening to the NATO?
> 2. Who sees this move as a threat?
> 3. Who is bearing the main cost?

Table 4: Evaluation of machine and human authored summaries using questions by the different subject.

Figure 3. Questions are set by the different author from the one who wrote the summary. A human authored summary may still be the best summary in many respects, but it will no longer be considered perfect. One may target the relevance level of the human summary (*e.g.*, 61% for the 'one line' summary task from the broadcast news stories) for automatic summarisation research.

Table 4 shows one example from those with which we are currently experimenting. Answers sought by Subject **D** were '*expansion*', '*Russian*', and '*American taxpayers*', respectively. Given this question set, answers are '*relevant*', '*relevant*', and '*not found*' for the summary by Subject **B**, and answers found in the machine generated summary are '*relevant*', '*not found*', and '*not found*', respectively.

7 Conclusion

In this paper, we have presented the issue of human subjectivity when authoring summaries, with regard to producing a simple, robust evaluation of machine generated summaries. Applying the cross comprehension test on human authored 'one line' summaries from broadcast news stories, we gauged the level of subjectivity among four authors. The instruction set was simple, thus there was enough room for subjectivity. However the approach was robust because the test did not use the absolute score, instead relying on relative comparison, effectively alleviating the subjectivity. We also showed the approach to evaluating machine generated summaries. The experiment using this scheme is currently underway.

References

C. Cieri, D. Graff, M. Liberman, N. Martey, and S. Strassel. 1999. The TDT-2 text and speech corpus. *DARPA Broadcast News Workshop*, Herndon, VA.

M. Hirohata, Y. Shinnaka, K. Iwano, and S. Furui. 2005. Sentence extraction-based presentation summarization techniques and evaluation metrics. *ICASSP*, Philadelphia.

L. Hirschmann, J. Burger, D. Palmer, and P. Robinson. 1999. Evaluating content extraction from audio source. *ESCA Workshop: Accessing Information in Spoken Audio*, Cambridge.

C. Hori, T. Hori, and S. Furui. 2003. Evaluation method for automatic speech summarization. *Eurospeech*, Geneva.

C. Lin and E. Hovy. 2003a. Automatic evaluation of summaries using n-gram co-occurrence statistics. *HLT-NAACL*, Edmonton.

C. Lin and E. Hovy. 2003b. The potential and limitations of automatic sentence extraction for summarization. *HLT-NAACL Workshop on Automatic Summarization*, Edmonton.

I. Mani. 2001. *Automatic Summarization*. Jon Benjamins Publishing Company.

A. Nenkova and R. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. *HLT-NAACL*, Boston.

P. Over and J. Yen. 2004. An introduction to DUC 2004: Intrinsic evaluation of generic news text summarization systems. *DUC Workshop*, Boston.

D. Radev and D. Tam. 2003. Summarization evaluation via relative utility. *CIKM*, New Orleans.

H. Van Halteren and S. Teufel. 2003. Examining the consensus between human summaries: Initial experiments with factoid analysis. *HLT-NAACL Workshop on Automatic Summarization*, Edmonton.

Appendix

Attached below is a complete news story for the human authored summaries in Section 3. It is taken from '*ABC News*', aired on 13 February 1998, between 1830 and 1900 hours, where Peter Jennings was in the studio, and John Mackenzie was reporting from Central Park:

"In case you had forgotten and probably you shouldn't have tomorrow is Valentine's day. It is largely an American celebration though other parts of the world are picking up on it. We're told whether we are married or single male or female. We know what we're supposed to do don't we? We thought tonight without getting too serious about it we would take a closer look at the science of love. I confess that we never thought of it as science until yesterday when we went across to the Central Park Zoo here in New York to meet and have a conversation with Dr Helen Fisher from Rutgers University in New Jersey. Dr Fisher is a noted anthropologist who has been studying the behaviour between men and women for many years."

"Dr Fisher can I ask you is this really serious science that you do?"

"I think it's serious yes. I'm interested in minding the brain physiology of very basic human mating emotion attraction and I think it comes out of nature."

"Can you break down for me what the components of attraction are?"

"It begins when a person takes on special meaning. Indeed you focus on that individual. There's another thing called intrusive thinking. The person pops into your brain. Some people have said I think about him or her eight five of the day. People also focus on their sweetheart. They will remember a tiny little thing that the person said or did. Just the way they toss their head when they got off the bus or reached for the salt at the dinner table. And then of course there's that elation and that giddiness and euphoria and that tremendous despair when the person doesn't call you. When men and women begin to fall in love do they do it differently. Men tend to fall in love faster. I think because men are more turned on by the visual image. A man can scan a room and see a woman who really appeals to him. The woman has to find out whether the man has he resources whether he's a find individual."

"What are the visual traits for a man when he's in this process?"

"The visual traits that a man will be attracted to will vary dramatically. We all grow up as small from childhood and we build an unconscious love map. A whole list of traits that we are individually looking for in a mate. For example our father's sense of humour. The amount of chaos around the house. Subtle little things will get into the brain and we will create almost a testimony plate of what we are looking for."

"Are different cultures attracted in different ways?"

"There are some ways in which people in every culture are attracted in the same way. Men around the world are attracted to women who give off signs of fertility. Clear skin bright eyes a great personality the kinds of things that indicate that a woman would be good at bearing his young. Women around the world are interested in men who have resources status class money the kinds of things that would help them rear their young. Around the world both men and women are attracted to a face that is symmetrical."

"Doesn't matter whether you are Asian or American?"

"No, you and I could go to New Guinea and you and I would be able to pick out what we regarded as the most beautiful woman in the village and the villagers would agree with us."

"What's the difference of the attraction being dominated by brain and culture?"

"I think human beings evolve certain circuits in the brain that light up when you see the right person. Those circuits are largely associated with dopamine chemicals in the brain that give you that sense of elation and giddiness and euphoria when you see the right person. That's the brain chemistry of romance."

"But who you fall in love with when you fall in love where you fall in love how you express your love that's cultural?"

"That you learn."