
Auto-Highlighter: Identifying Salient Sentences in Text

Jessica Zeitz Self
Rebecca Zeitz
Alan Breitler
Chris North

Motivation

Highlight Visualizer

An Information Saliency Tool

Time Wars (CS)

Mark Fisher

Time rather than money is the currency in the recent science fiction film *In Time*. At the age of 25, the citizens in the future world the film depicts are given only a year more to live. To survive any longer, they must earn extra time. The decadent rich have centuries of empty time available to fritter away, while the poor are only days or hours away from death. *In Time* is, in effect, the first science fiction film about precarity – a condition that describes an existential predicament as much as it refers to a particular way of organising work.

At the most simple level, precarity is one consequence of the "post-Fordist" restructuring of work that began in the late 1970s: the turn away from fixed, permanent jobs to ways of working that are increasingly casualised. Yet even those within relatively stable forms of employment are not immune from precarity. Many workers now have to periodically revalidate their status via systems of "continuous professional development"; almost all work, no matter how menial, involves self-surveillance systems in which the worker is required to assess their own performance. Pay is increasingly correlated to output, albeit an output that is no longer easily measurable in material terms.

For most workers, there is no such thing as the long term. As sociologist Richard Sennett put it in his book *The Corrosion of Character: The Personal Consequences of Work in the New Capitalism*, the post-Fordist worker "lives in a world marked ... by short-term flexibility and flux ... Corporations break up or join together, jobs appear and disappear, as events lacking connection." (30) Throughout history, humans have learned to come to terms with the traumatic upheavals caused by war or natural disasters, but "[w]hat's peculiar about uncertainty today," Sennett points out, "is that it exists without any looming historical disaster; instead it is woven into the everyday practices of a vigorous capitalism."

It is not only that we are incapable of even conceiving of different ways to work, produce and consume. It's now clear that, from the beginning (and with good reason) neoliberalism declared war on this alternative mode of time. It remains tiresome to see the propagation of resentment against those few fugitives who can still escape the treadmill of debt and work, promising to ensure that soon, they too will be condemned to performing interminable, meaningless labour – as if the solution to the current stagnation lay in more work, rather than an escape from the work. If there is to be any kind of future, it will depend on our winning back the uses of time that neoliberalism has sought to close off and make us forget.

Highlight Visualizer

An Information Saliency Tool

Time Wars (non-CS)

Mark Fisher

Time rather than money is the currency in the recent science fiction film *In Time*. At the age of 25, the citizens in the future world the film depicts are given only a year more to live. To survive any longer, they must earn extra time. The decadent rich have centuries of empty time available to fritter away, while the poor are always only days or hours away from death. *In Time* is, in effect, the first science fiction film about precarity – a condition that describes an existential predicament as much as it refers to a particular way of organising work.

At the most simple level, precarity is one consequence of the "post-Fordist" restructuring of work that began in the late 1970s: the turn away from fixed, permanent jobs to ways of working that are increasingly casualised. Yet even those within relatively stable forms of employment are not immune from precarity. Many workers now have to periodically revalidate their status via systems of "continuous professional development"; almost all work, no matter how menial, involves self-surveillance systems in which the worker is required to assess their own performance. Pay is increasingly correlated to output, albeit an output that is no longer easily measurable in material terms.

For most workers, there is no such thing as the long term. As sociologist Richard Sennett put it in his book *The Corrosion of Character: The Personal Consequences of Work in the New Capitalism*, the post-Fordist worker "lives in a world marked ... by short-term flexibility and flux ... Corporations break up or join together, jobs appear and disappear, as events lacking connection." (30) Throughout history, humans have learned to come to terms with the traumatic upheavals caused by war or natural disasters, but "[w]hat's peculiar about uncertainty today," Sennett points out, "is that it exists without any looming historical disaster; instead it is woven into the everyday practices of a vigorous capitalism."

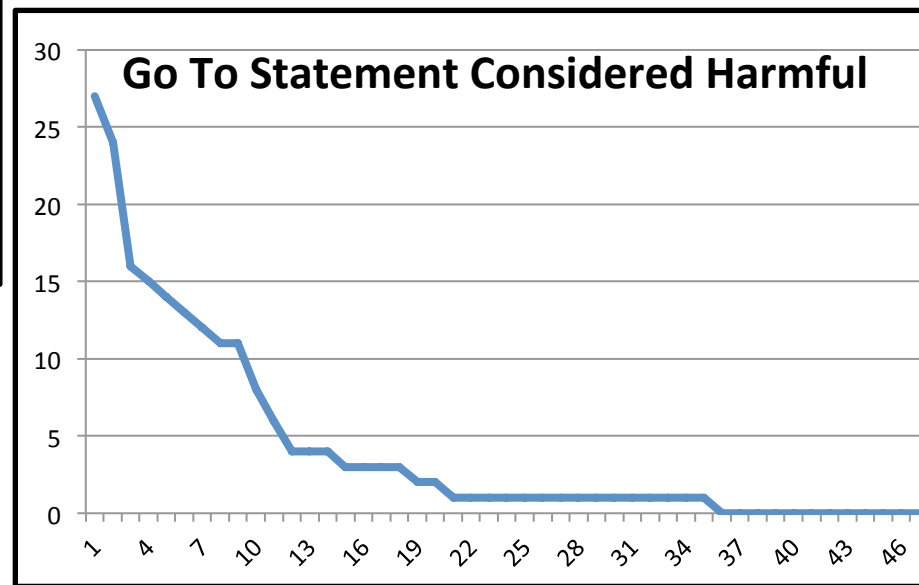
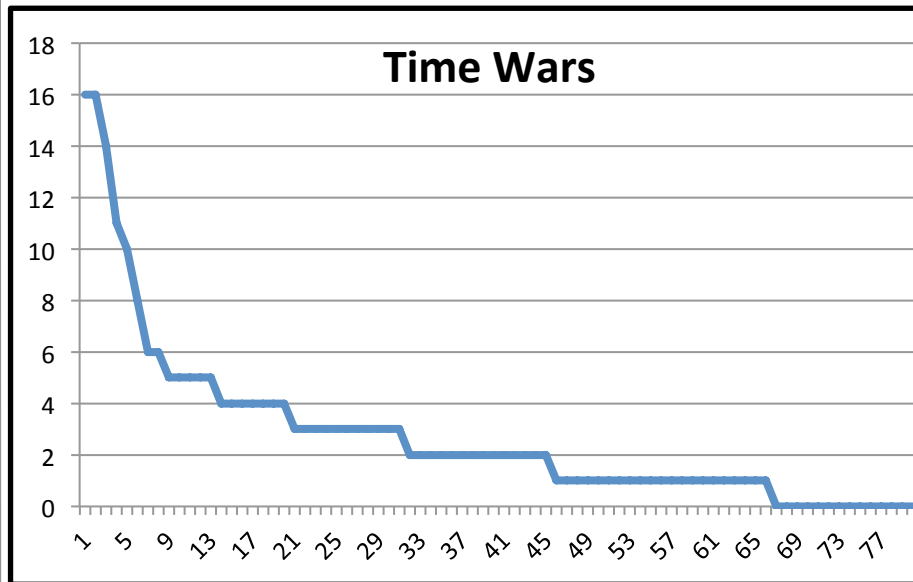
Research Questions

1. Are human highlighted sentences representative of human synthesized summaries? When summarizing, how much do humans rely on sentences they previously highlighted?
2. Which sentences in a document do humans deem salient? Why?
3. Are there differences between experts and non-experts when highlighting and summarizing a document?
4. How closely can simple algorithm heuristics mimic the human selection of salient sentences?

The Experiment

- Independent variables
 - document type
 - participants' area of expertise
- 40 participants
- 2 documents (technical and non-technical)
- Dependent variables
 - highlights
 - summaries
 - reasons for highlights
 - relation between highlights and summaries

The Findings



Characterization of Summaries (Q1)

Connecting Highlights and Summaries

Reference Location	Go To Statement		Time Wars	
	<i>Non-technical</i>	<i>Technical</i>	<i>Non-technical</i>	<i>Technical</i>
Highlighted sentences	51%	50%	44%	59%
Elsewhere in document	10%	8%	41%	32%
Synthesized information	39%	42%	15%	9%

Highlights are 50% representative of summaries

User-defined Categories (Q2)

Prominent Categories of Reasons for Highlights

Category	Go To Statement		Time Wars	
	<i>Non-technical</i>	<i>Technical</i>	<i>Non-technical</i>	<i>Technical</i>
Argument/ main point	13	27	14	11
Supporting evidence	54	45	54	54
Solution	7	0	7	1
Profound statement	n/a	n/a	6	0
Personally resonated	n/a	n/a	0	6
Conclusion	10	6	4	8

Rhetorical Structure (Q2)

- Strong correlation between categories and rhetorical elements
- Higher prevalence of formalized elements
 - main point
 - supporting evidence
- Categories fit order of rhetorical elements

Experts versus Non-experts (Q3)

- Minimal variations
- Correlation = 0.82
- Followed same overall trend
- Employed same methods

User and Algorithm Comparison (Q4)

- Metric based on uniqueness performed poorly
- Small correlation with sentence length
- Metric that found sentences representative of whole text performed well

Correlation R Values between Metrics & Humans

Metric	Go To Statement	Time Wars
# of n-grams	0.542	0.288
sentence length	0.335	0.046
tf-idf	0.612	0.398

Comparison Visual

Algorithm

Highlight Visualizer

An Information Saliency Tool

Go To Statement Considered Harmful (Algorithm)

Edsger Dijkstra

For a number of years I have been familiar with the observation that the quality of programmers is a de function of the density of go to statements in the programs they produce. More recently I discovered w use of the go to statement has such disastrous effects, and I became convinced that the go to stateme should be abolished from all "higher level" programming languages (i.e. everything except, perhaps, pla machine code). At that time I did not attach too much importance to this discovery; I now submit my considerations for publication because in very recent discussions in which the subject turned up, I have urged to do so.

My first remark is that, although the programmer's activity ends when he has constructed a correct pro the process taking place under control of his program is the true subject matter of his activity, for it is process that has to accomplish the desired effect; it is this process that in its dynamic behavior has to the desired specifications. Yet, once the program has been made, the "making" of the corresponding pr delegated to the machine.

My second remark is that our intellectual powers are rather geared to master static relations and that o powers to visualize processes evolving in time are relatively poorly developed. For that reason we shou (as wise programmers aware of our limitations) our utmost to shorten the conceptual gap between the program and the dynamic process, to make the correspondence between the program (spread out in tex space) and the process (spread out in time) as trivial as possible.

Humans

Highlight Visualizer

An Information Saliency Tool

Go To Statement Considered Harmful (Humans)

Edsger Dijkstra

For a number of years I have been familiar with the observation that the quality of programmers is a de function of the density of go to statements in the programs they produce. More recently I discovered w use of the go to statement has such disastrous effects, and I became convinced that the go to stateme should be abolished from all "higher level" programming languages (i.e. everything except, perhaps, pla machine code). At that time I did not attach too much importance to this discovery; I now submit my considerations for publication because in very recent discussions in which the subject turned up, I have urged to do so.

My first remark is that, although the programmer's activity ends when he has constructed a correct pro the process taking place under control of his program is the true subject matter of his activity, for it is process that has to accomplish the desired effect; it is this process that in its dynamic behavior has to the desired specifications. Yet, once the program has been made, the "making" of the corresponding pr delegated to the machine.

My second remark is that our intellectual powers are rather geared to master static relations and that o powers to visualize processes evolving in time are relatively poorly developed. For that reason we shou (as wise programmers aware of our limitations) our utmost to shorten the conceptual gap between the program and the dynamic process, to make the correspondence between the program (spread out in tex space) and the process (spread out in time) as trivial as possible.

In Conclusion

Highlights

- ... are 50% representative of human summaries.
- ... do fit within broad categories.
- ... do correlate with rhetorical structure.
- ... are chosen by experts and non-experts using the same methods.
- ... chosen by an algorithm using simple text metrics are representative of human summaries.

References

1. G. J. Rath, A. Resnick, and T. . Savage, “The Formation of Abstracts By the Selection of Sentences Part I. Sentence Selection By Men and Machines,” *Journal of the American Society for Information Science and Technology*, vol. 12, no. 2, pp. 139–141, 1961.
2. D. Radev, S. Teufel, and H. Saggion, “Evaluation challenges in large- scale document summarization,” *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, vol. 1, pp. 375– 382, 2003.
3. E. W. Dijkstra, “Letters to the editor: go to statement considered harmful,” *Communications of the ACM*, vol. 11, no. 3, pp. 147–148, Mar. 1968.
4. M. Fisher, “Time Wars,” *Gonzo (circus)*, no. 110, 2012.
5. X. Wang and A. Kabán, “Model-based estimation of word saliency in text,” *Discovery Science*, 2006.