A Hard Look at Soft Concepts

Dafna Shahaf
The Hebrew University of Jerusalem
dshahaf@cs.huji.ac.il

Abstract

What can *only* humans do? What can humans do that machines cannot? These questions have long been tantalizing scientists and philosophers. Many areas, such as creativity and humor, are traditionally considered to be outside the reach of computers. We believe that these territories define an intriguing set of challenges for computer science. We present two approaches for tackling such challenges – an axiomatic one and a data-driven one – and demonstrate our ideas on two real-world applications: finding narratives in large textual corpora and identifying humorous cartoon captions.

1 Introduction

Daniel Gilbert says that all psychologists must, at some point in their professional lives, publish an article containing this sentence: "The human being is the only animal that..." [Gilbert, 2009]. Indeed, many versions of this sentence have been written over the years; often, they are later discovered to be wrong and are discarded.

Initially, those sentences were debunked by studies of animals. For example, humans were considered to be the only animal that could use language, until we found that chimpanzees could learn sign language. Today, the battle for humans' sense of uniqueness is increasingly waged against *computers*. Computers can now perform many tasks that were once considered uniquely human, such as playing Go [Silver *et al.*, 2016] or Jeopardy [Ferrucci *et al.*, 2010].

Today, there are still many areas that are considered to be outside the reach of computer science. For example, creativity and humor are viewed as distinctly human traits; even in science fiction, robots and computers are almost always portrayed as humorless – no matter how proficient they are with language or other skills.

We believe that such territories define an intriguing set of challenges for computer science. Improvements in these areas can immediately translate to improvements in human-computer interaction and collaboration; even more importantly, they could lead to new insights about human cognition. In the following, we present two approaches for tackling such challenges. The **axiomatic** approach aims to formalize intuitive concepts, crafting an objective function to emulate elusive intuitions. We demonstrate this approach in Section 2, summarizing methods we have developed to capture the structure and development of complex news stories.

Coming up with objective functions can be difficult. Luckily, recent advancements in data collection and machine learning have made it easier to discover potential axioms. The **data-driven** approach uses data to guide the objective formulation. In Section 3, we show how we used crowdsourced data to automatically identify humorous cartoon captions.

2 Axiomatic Approach: Information Cartography

We now demonstrate the axiomatic approach, summarizing methods we have previously developed to tackle information overload [Shahaf *et al.*, 2012b; 2012a; 2013; 2015a].

When information is abundant, people struggle to make sense of complex issues, such as presidential elections or economic reforms. Methods that summarize and visualize narratives [Swan and Jensen, 2000; Yan *et al.*, 2011; Allan *et al.*, 2001] often work only for simple (and linear) stories. In contrast, complex stories are non-linear: stories spaghetti into branches, side stories, and intertwining narratives. To explore such stories, users need a *map* to guide them through unfamiliar territory.

We have introduced a methodology for creating structured summaries of information, which we call *metro maps*. Metro maps consist of a set of lines which can intersect or overlap. Most importantly, metro maps explicitly show the relations among different pieces in a way that captures the evolution of a story. Each metro stop is a cluster of articles, and lines follow coherent narrative threads. Different lines focus on different aspects of the story.

For example, Figure 1 summarizes the 2014 Crimean crisis. The map was automatically generated for the query "Crimea". The lines correspond to the Russian, Ukrainian, and Western points of view. Timeline appears at the bottom, and important words appear next to each line. The Russian (green) line starts with the Crimea parliament voting and Putin recognizing Crimean independence. The Ukrainian (orange) line starts with Ukraine's former prime minister urging the West to stop Russian aggression. The Ukrainian line then joins the Western (blue) line, discussing the West's attempts to support Ukraine. The Russian and Ukrainian lines intersect

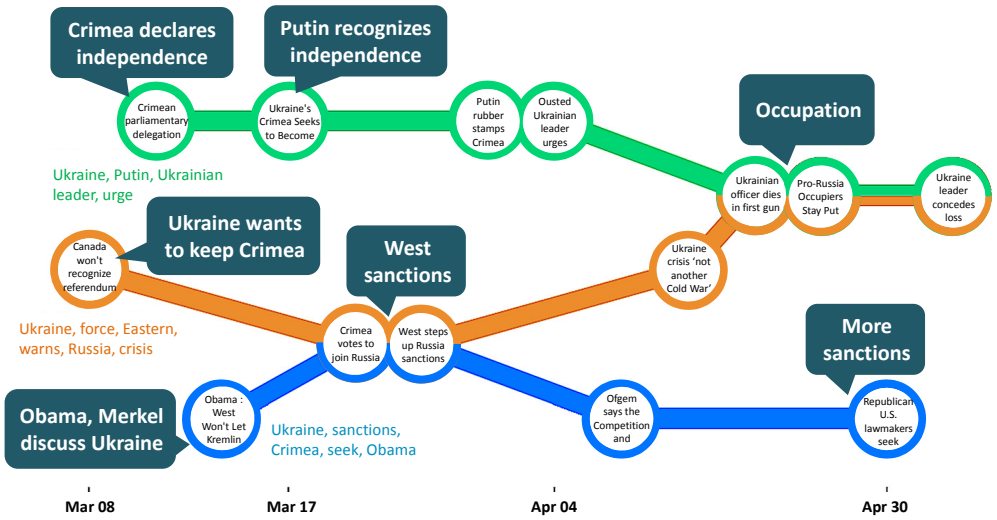

Figure 1: Sample output: Metro map about the 2014 Crimean crisis. Legend on the left of each line shows the important words for the line: the lines correspond to the Russian, Ukrainian, and Western points of view. Each metro stop is a cluster of articles (the callout bubbles are manual annotations of the content). The timeline appears at the bottom of the map.

when pro-Russia men take over police stations in Ukraine.

2.1 Crafting an Objective Function

Before we can come up with an algorithm for computing good maps, we must craft an objective function. This is especially important because the objective is not a priori clear: Recognizing whether a map is good or not is easy for humans, but it is a very intuitive notion.

In a nutshell, the axiomatic approach attempts to formalize soft, intuitive concepts by breaking them down into a list of axioms. For example, later we argue that the map objective must guarantee *diversity* of articles, and that guarantee should be formalized as a submodular function over sets of articles.

Let us recall our goal again: We seek to compute a *metro map* to summarize and organize an input set of documents. Metro lines are ordered sequences of stops (which are subsets of articles). Line follow coherent narrative threads, with different lines focus on different aspects. Intersections reveal interactions between storylines. This intuitive definition gives rise to a set of (sometimes conflicting) criteria, such as *coherence* of lines. In the following, we motivate and formalize these criteria, resulting in a set of axioms we need to satisfy.

Coherence. A first requirement is that each metro line tells a *coherent* story: Following the articles along a line should give the user a clear understanding of the evolution of a story.

For the sake of the presentation, we focus metro stops that are singletons: each cluster is a single document. A natural first step in defining coherence is to measure similarity between each pair of consecutive documents. However, local similarities may give rise to associative, incoherent lines. For example, suppose a line consists of three documents. Documents 1,2 could share half their words, as well as 2,3, but documents 1,3 may still have nothing in common.

We note that coherent lines are often characterized by a *small* set of words that are important throughout the line. In other words, coherence is a *global* property of the line, and cannot be deduced solely from local interactions.

We translate the axioms into a linear programming problem, where the goal is to choose a small set of words, and score the line based solely on them. To ensure that each transition is strong, the score of a chain is the score of the weakest link. See [Shahaf and Guestrin, 2010] for details.

Coverage. *Coherence* is crucial for good maps, but it is not sufficient. Maximally coherent lines often revolve around unimportant, esoteric topics. Furthermore, as there was no notion of diversity, multiple lines contained redundant information. Thus, another key property is *coverage*: lines should cover diverse topics which are important to the user.

We define a set of elements we wish to cover: For news articles, we use words (e.g., "Obama", "China") [Shahaf *et al.*, 2012b], so a high-coverage map discusses many important words. For the scientific corpus, we use papers [Shahaf *et al.*, 2012a], so map should cover large part of the corpus.

We define a coverage function, measuring how well a set of documents covers each element. To encourage diversity, this function should be *submodular*. Thus, if the map already covers an element well, similar documents provide little extra coverage, encouraging us to cover new topics. In addition, *weights* biases the map towards covering important elements. In [Shahaf *et al.*, 2012b], we discuss learning weights from user feedback for a personalized notion of coverage.

Connectivity. Finally, a map is more than just its content; there is information in its *structure* as well. Our last property is *connectivity*: The map should convey the underlying structure of the story, and how aspects of the story interact.

Intuitively, different stories have different structure. Some stories are almost linear, while others are much more complex. In order to capture the structure of a story, we compute the minimum number of lines that cover all metro stops. This

objective pushes for long storylines whenever possible: linear stories become linear maps, while complex stories maintain their interweaving threads.

Tying it all Together. Next, we consider tradeoffs among the properties we defined. For example, maximizing coherence often results in low coverage. Therefore, it is better to treat coherence as a constraint: a chain is either coherent enough to be included in the map, or not. Coverage and connectivity, on the other hand, should both be optimized. Finally, we define our objective:

Problem 2.1 (Metro Maps: Informal)**.**
A map should satisfy:

- *High coverage, High connectivity*

Subject to:

- *Minimal level of line coherence, Minimal cluster quality, Maximal map size*

See [Shahaf *et al.*, 2013] for a formal statement of the axioms, as well as the optimization algorithm.

2.2 Applications and Evaluation

Maps are very easy to apply to domains other than news. The principles stay the same, but one can use domain knowledge to adapt the objectives (while still obeying the axioms). In addition to news, we applied the algorithm to science (help researchers understand the state-of-the-art), legal documents (help lawyers prepare for a case), and books (elucidate the structure of complex books – in particular, Lord of the Rings).

Evaluating metro maps is difficult, as ground truth is hard to define. Since the goal of the maps is to help people navigate through information, we conducted an extensive set of user studies to better understand the value of the methodology. For lack of space, we only describe one study in the scientific domain. In that study, we recruited students and asked them to conduct a literature survey in an area they were not familiar with. We measured *precision* (fraction of retrieved papers that are relevant), and *subtopic recall* (fraction of relevant research areas retrieved). Map users outperformed Google Scholar users in every parameter: Precision (84.5% to 74.2%), recall (73.1% to 46.4%) and number of seminal papers found (1.62 to 1.2).

3 Data-Driven Approach: Identifying Humorous Cartoon Captions

In the previous section we demonstrated the axiomatic method. We now demonstrate the **data-driven** method in the humor domain. A considerable amount of discussion has been devoted to the nature and function of humor. To date, most investigations of humor have been undertaken within psychology, philosophy, and linguistics, while computational research is still in its infancy.

We believe that endowing machines with capabilities for recognizing humor could enhance interaction through improved understandings of semantics, intentions, and sentiment. Humor can also be harnessed to increase attention, retention, and engagement, and thus has numerous applications in education, health, communications, and advertising.

What's it going to take to get you in this car today?
Relax! It just smells the other car on you.
It runs entirely on legs.
Just don't tailgate during mating season.
It's only been driven once.
He even cleans up his road kill.
The spare leg is in the trunk.
Comfortably eats six.
She runs like a dream I once had.

Figure 2: Example cartoon from the New Yorker contest, with the shortlist of submitted captions.

Computational work to date on humor focuses largely on humor generation in limited domains, such as puns and humorous acronyms [Binsted and Ritchie, 1994; Stock and Strapparava, 2005]. Several other projects focus on the related task of humor recognition [Mihalcea and Pulman, 2007; Taylor and Mazlack, 2004; Tsur *et al.*, 2010].

In this section we describe a new direction we proposed in the realm of humor recognition [Shahaf *et al.*, 2015b]. We focus on the task of identifying humorous captions for cartoons. Specifically, we consider cartoons from The New Yorker magazine. The New Yorker holds a weekly contest in which it publishes a cartoon in need of a caption. Readers are invited to submit their suggestions for captions. The judge selects a shortlist of the funniest captions, and members of the editorial staff narrow it down to three finalists. All three finalists' captions are then published in a later issue, and readers vote for their favorites. Figure 2 shows an example cartoon. In this cartoon, a car salesperson attempts to sell a strange, hybrid creature that appears to be part car, part animal. The judge's shortlist appears under the cartoon.

3.1 Crafting an Objective Function

We wish to formulate the humorousness of the captions. Unlike the previous section, it is much harder to describe desired properties of funny captions. This task seems to require understanding deeper, more universal underpinnings of humor.

Thus, we use data to guide our process. Unlike the axiomatic approach, in the data-driven approach we do not state in advance desired properties we must satisfy. Rather, we construct candidate axioms from data – either from the digital traces we all leave behind or through direct experimentation.

For our task, we create a dataset of crowdsourced New Yorker competition captions, along with human judgments. We recruited crowdworkers via Mechanical Turk. Workers saw a cartoon and multiple captions, and proceeded to rank the captions from the funniest to the least funny. We selected

pairs of captions that achieved high agreement among the rankers (80% agreement or more, similar length, ranked by at least five people). These pairs served as our ground truth.

Next, we conducted a search of the humor-research literature, looking for useful attributes for discriminating funnier captions. For example, Mihalcea and Pulman [Mihalcea and Pulman, 2007] suggest a frequent use of negative words (cannot, bad, illegal, mistake) in humorous texts. In this spirit, we used a sentiment analysis tool to annotate the captions. Patrick House, a winner of the New Yorker contest, wrote an article about his winning strategy [House, 2008]. He suggests to use "common, simple, monosyllabic words" and to "Steer clear of proper nouns that could potentially alienate". Following this advice, we measure readability metrics and proper nouns. Similarly, we extracted multiple other potential hypotheses from the literature (some of them contradicting), as well as some of our own – in particular, metrics that take the cartoon itself into account.

Using our data, we could test potential axioms. For example, we found that funnier captions indeed use significantly fewer proper nouns and 3rd person words. This information enabled us to try and formalize a notion of humor. See details in [Shahaf *et al.*, 2015b].

3.2 Evaluation

We formulate a pairwise evaluation task and construct a classifier that, given two captions and a cartoon, determines which caption is funnier. Our classifier achieves 69% accuracy for captions hinging on the same joke, and 64% accuracy comparing any two captions. We implemented an algorithm that ranks all captions. On average, all of the judges' top-10 captions are ranked in the top 55.8%, thereby suggesting that the methods can be used to significantly reduce the workload faced by judges.

4 Conclusions

We proposed two ways to tackle areas that are traditionally considered to be outside the scope of computers. We used an axiomatic approach to construct metro maps, concise structured sets of documents that follow the development of complex stories. We formalized soft terms like "coherence" and "coverage", resulting in objectives that capture people's intuitive notions. We applied metro maps to help people understand news stories, research areas, legal cases, and works of literature. User studies suggest that metro maps help users to acquire knowledge efficiently.

Next, we demonstrated a data-driven approach in the humor domain. We investigated the challenge of learning to recognize the degree of humor perceived for cartoons and captions. We extracted a useful set of features from linguistic properties of captions and the interplay between captions and pictures. We harnessed a large corpus of crowdsourced cartoon captions and developed a classifier that could pick the funnier of two captions 64% of the time. We used the classifier to find the best captions, significantly reducing the load on the cartoon contest's judges.

We believe that formalizing areas that are viewed as "distinctively human" will find myriad applications across multiple domains. Beyond applications, pursuing formal models could reveal new insights about fascinating aspects of the human cognition.

Acknowledgments. The author would like to thank Carlos Guestrin, Eric Horvitz, Jure Leskovec and Bob Mankoff. The author is a Harry&Abe Sherman assistant professor, and is supported by ISF grant 1764/15.

References

[Allan *et al.*, 2001] J Allan, R Gupta, and V Khandelwal. Temporal summaries of new topics. In *SIGIR'01*. ACM, 2001.

[Binsted and Ritchie, 1994] K Binsted and G Ritchie. An implemented model of punning riddles. In *AAAI'94*, 1994.

[Ferrucci *et al.*, 2010] D Ferrucci, E Brown, J Chu-Carroll, et al. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010.

[Gilbert, 2009] D Gilbert. *Stumbling on happiness*. Vintage Canada, 2009.

[House, 2008] P. House. How to win the new yorker cartoon caption contest. Slate, 2008.

[Mihalcea and Pulman, 2007] R Mihalcea and S. Pulman. Characterizing humour: An exploration of features in humorous texts. In *Computational Linguistics and Intelligent Text Processing*. Springer, 2007.

[Shahaf and Guestrin, 2010] D Shahaf and C Guestrin. Connecting the dots between news articles. In *KDD'10*, 2010.

[Shahaf *et al.*, 2012a] D Shahaf, C Guestrin, and E Horvitz. Metro maps of science. In *KDD'12*. ACM, 2012.

[Shahaf *et al.*, 2012b] D Shahaf, C Guestrin, and E Horvitz. Trains of thought: Generating information maps. In *WWW'12*, pages 899–908. ACM, 2012.

[Shahaf *et al.*, 2013] D Shahaf, J Yang, C Suen, J Jacobs, H. Wang, and J Leskovec. Information cartography: Creating zoomable, large-scale maps of information. In *KDD'13*. ACM, 2013.

[Shahaf *et al.*, 2015a] D Shahaf, C Guestrin, E Horvitz, and J Leskovec. Information cartography. *Commun. ACM*, 58(11):62–73, October 2015.

[Shahaf *et al.*, 2015b] D Shahaf, E Horvitz, and R Mankoff. Inside jokes: Identifying humorous cartoon captions. In *KDD'15*, pages 1065–1074, New York, NY, USA, 2015. ACM.

[Silver *et al.*, 2016] D Silver, A Huang, CJ Maddison, A Guez, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[Stock and Strapparava, 2005] O Stock and C Strapparava. Hahacronym: A computational humor system. In *Proceedings of the Association for Computational Linguistics*. ACL, 2005.

[Swan and Jensen, 2000] R Swan and D Jensen. TimeMines: Constructing Timelines with Statistical Models of Word Usage. In *KDD'00*, 2000.

[Taylor and Mazlack, 2004] J Taylor and L Mazlack. Computationally recognizing wordplay in jokes. *CogSci'04*, 2004.

[Tsur *et al.*, 2010] O Tsur, D Davidov, and A Rappoport. Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *ICWSM*, 2010.

[Yan *et al.*, 2011] R Yan, X Wan, J Otterbacher, L Kong, X Li, and Y Zhang. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *SIGIR'11*, 2011.