PinfoViz: a browser extension for annotating numeric mentions with personalized analogies and visualizations

Tum Chaturapruek
Computer Science Department
Stanford, CA
sorathan@cs.stanford.edu

ABSTRACT

Distances and numerical measures are common in text articles. Yet these numbers, while precise, are difficult to comprehend when readers cannot relate the numbers to personal experience. Since common knowledge is a relative concept, other solution frameworks that do not take personal information into account suffer from this problem. We develop a framework and a Chrome extension that leverage personal information to contextualize numbers inline as readers read numbers in a browser, where personal information is collected either from a social media account login, manual inputs, or persona template selections.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords

Distance; personalization; spatial analogy; information visualization; landmark.

INTRODUCTION

Nowadays writing and reading are indispensable mode of communicating new information in the modern world. Distances and numerical measures are common in text and news articles. Yet these numbers, while precise, are difficult to comprehend when readers cannot relate the numbers to personal experience. Unclear communication result in false impressions, which can lead people to later on make uninformed decisions. This issue not only a problem for the numerically illiterate; even numerically literate people are struggling to interpret numbers that are out of the context or unfamiliar.

For example, take a look at Figure 1. We see 180 miles north of Alexandria, Egypt in a CNN news article. To me, a non-driver, I do not have a clear sense of how far this is. I also do not want to leave the site, open a Google Map and search for a place that is north of Alexandria by 180 miles. Even if I were successful in finding such a place, I could still have no idea what 180 miles really means.

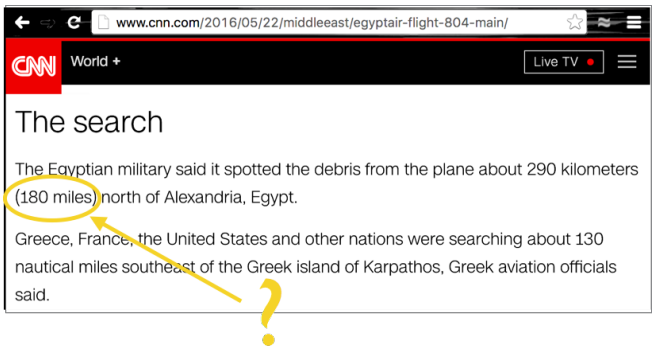

Figure 1. 180 miles in a CNN news article is difficult to comprehend, especially for readers who do not often travel by car.

The challenges are that 1) Finding a helpful analogy to contextualize is difficult because "common" knowledge is relative, and 2) Requiring users to go to another site to contextualize these numbers induces a big usage barrier.

To address this problem, we realize that personal information is not only nice to have, but *is necessary*. To see this, imagine that we do not have personal information about the user other than user location. Then effectively we lump all users in that location into one persona, while in fact people have very different walks of lives and experience, and also very different ways of thinking and perceive information. With this thought in mind, we propose a solution that directly takes personal information into account.

Contributions

- We present a novel solution framework that is end-to-end from personal information gathering, to how to incorporate this information into the energy scoring function, and to how to design an effective user interface that minimizes the usage barrier.
- We make a Chrome browser extension that users can easily install and use.
- We test our system with users. Initial qualitative feedback on the design is positive in that this system can help users better understand numbers in text articles. We need to finish implementing more features before we can fully test the end-to-end system.
- We generate ideas on how to learn from this tool about what information different person types find understandable, on

how writers can leverage this information to write a clearer and more personalized piece of writing, and on myriad ways to contextualize numbers.

RELATED WORK

Kim et al. [5] solve the same task of automatically generating spatial analogies among well-known landmarks near user's location. They also demonstrate that the system statistically significantly improves the helpfulness rating of information understanding. This system personalizes spatial analogies for users by taking user location into account. However, all users who stay at the given location are not equally familiar (as compared to their neighbors) with the landmarks around them. Thus, in this paper we attempt to leverage more personal information other than locations to resolve this issue. For examples, we factor places that users have visited into our energy scoring function. The user interface design choice between our work and that of Kim et al. is also different: we draw a circle around the user location to so that we can ensure a reasonable and reliable baseline, and we include multiple landmarks into the map so that users have more options to compare with. The relative effectiveness of the two interface designs is a topic for future work.

Wolfram|Alpha [2] can convert 10,000 miles to "*1/43 solar radius*." Another notable feature is that it presents multiple senses of the length comparison such as comparison as length and comparison as circumference. Our approach is different in that it is personalized and presents a lower usage barrier because users do not need to go to another site or pull information. Furthermore, since we have full information about the content of the page that users are visiting and not just a little number snippet that users enter into Wolfram|Alpha, in our framework it is therefore possible to draw context to disambiguate these multiple senses.

Another similar work is Dictionary of Numbers (Glen Chiacchieri). It can annotate "*315 million people*" as "*315 million people [≈ the population of the United States]*." This information is especially helpful if I live in the United States. If I live in China, this number might not be as helpful to me. Our approach aims to bridge this gap by leveraging personal information.

Chaganty and Liang [4] focus on compositing existing facts to make sense of the number. From its title, "*How Much is 131 Million Dollars?*", the paper describes that an annotation that the system generates to answer this question is "*about the cost to employ everyone in Texas over a lunch period.*" While this explanation does give a sense that the number is very large, it is still difficult *for me* to parse because I am not so familiar with the wage in Texas and the Texas population.

METHODOLOGY

Our high-level approach is the following:

- Collect and leverage personal information to contextualize numbers
- Factor personal information into the energy scoring function that ranks well-known landmarks to show to users
- Lower the usage barrier with inline annotations and tooltips in a browser

To address all above points, we choose to develop a Chrome browser extension. We will walk through our end-to-end pipeline from personal information collection to design considerations for user interface.

Approach to gather personal information

We design three methods to gather personal information:

- login with Facebook,
- manually input data, or
- choose from persona templates.

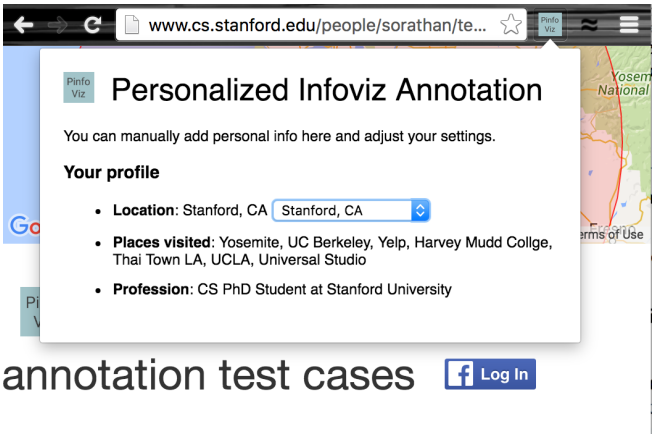

Figure 2. Personal information gathering interface.

Due to the time constraint, currently the only available input channel when clicking the extension icon is through a location template selection (e.g. "Stanford, CA"), see Figure 2. The places visited are hard-coded in the code. The facebook login feature requires passing a security requirement of an extension, and is currently off-loaded to a web site for this extension `http://www.cs.stanford.edu/people/sorathan/test_map.html`. Facebook automatically allows developers to request basic permissions from users: name, email, public profile, and user friends. Additional permissions require a review process. I have made an extension walk-through video and submitted it to request the following permissions: user tagged places, user work history, and user location. The requests are in review and normally take a week or two. Such information will be very helpful in personalizing annotations.

Finally, choosing from persona templates is a work in progress, but imagine that users can select that they are "a student at Stanford who likes outdoor activities", or "a Jazz music lover who also likes to visit art museums" or other persona in just one click. We also plan to allow users to further refine their settings through manual inputs. Note that to reason about which landmarks to show once we have the persona, we can cluster existing users and find a succinct description like above, and determine which kinds of landmarks they tend to be familiar with.

Approach to generate spatial analogies

1. *Find distance mentions in text articles*

   Since distances have expected forms, we use a regular expression to find the mentions.

   Specifically, we use `/([0-9]+)(,[0-9]+)*(.[0-9]+)?((hundred|thousand|million|billion|trillion))?(miles|mi|kilometers|km)/i`

2. *Use tooltips to show a Google Map and an explanation*

   The reason behind the design choice of tooltips is to present information to users in a non-invasive way.

   This step is surprisingly one of the most challenging steps among all development tasks (I have no prior experience developing a Chrome browser extension). The challenging part was how to properly set up callback functions from Google Maps APIs [1] to a function that exists only in the extension context ("content scripts"). Another layer of difficulty is to set up the right callback dependencies when we display tooltips. For tooltips, we use TippedJS [3], a javascript library for tooltips.

3. *Draw a radius from where the user is*

   The purpose of drawing a shaded circle with the radius equal to the target distance is to ensure that we have a reliable base line. In case the user does not have any familiar places near them, we are still able to show information that is useful to them.

4. *Find nearby well-known landmarks that the user has visited or is likely to be familiar with*

   Now that we have personal information, we can adjust the energy scoring function to take familiarity into account.

   Knowledge base generation
   This sub-problem is currently not our focus, since the contributions from Kim et al. [5] cover this problem. We simply take the final list and associate scores, and then modify those scores to further personalize to users.

   Energy scoring function
   Let u denote a user and l denote a landmark. Let $\mathscr{K}$ be the set of landmark candidates. Let $\hat{E}$ be the energy scoring function that summarizes which landmarks we list to users. We will include in the map those landmarks with high scores and exclude otherwise. We define

$$\hat{E}_u(l) = E_u(l) \times \frac{f_u(l)}{\sum_{k \in \mathscr{K}} f_u(k)},$$

   where $E_u(l) = E(u,l)$ is the same function as in [5], and f_u is the additional familiarity factor for each user to that place, a score that is only possible to compute in our system. Note that the energy function E already takes into account the distance between the user and the landmark, the general familiarity of the landmark, and different penalty scores for the multiplicative factors used to convert from the specific distance to the target distance (near 1.0 is optimal, too small multiplicative factor has a high penalty, and similarly too large a factor also has a high penalty).

   We introduce the new factor, f_u which is the additionnal familiarity factor for each user to that place. It will not be too dependent on the distance since that is already in E_u. We therefore use the following formula:

$$f_u(l) = \begin{cases} 1 & \text{if there is no relationship between user and landmark} \\ 1+n & \text{if user has never been to landmark, but been to places of the same category or vicinity } n \text{ times} \\ 1+3n & \text{if user has been to landmark } n \text{ times.} \end{cases}$$

   If users have no personal information inputs other than user location, note that the modified energy function $\hat{E}$ gracefully falls back to the normal energy function E.

5. *Use scaling to expand landmark candidates*

   Given a certain distance, it is possible that there are no landmarks that users are familiar with. We use the same approach as [5] to allow scaling. For example, if the target distance is 270 miles, we could say that it is 1.5 times the distance from Stanford to Yosemite.

   Note that very big or very small scaling factors can make the numbers not comprehensible (e.g. 10,000 miles is 1/43 of the solar radius, according to Wolfram|Alpha). Thus, we plan to limit the scaling factor to be an integral number or half an integral number and be in the range [0.5, 10].

   Note that this subtask has partially been done in the previous step (see [5]), but we impose more penalties on decimal numbers. By combining this feature, it allows us to avoid sparsity that other tools suffer from, while do not introduce unnecessary complexities.

Chrome extension code can be found at `https://bitbucket.org/sorathan/pinfo_viz` (please email the author to request access).

Design considerations

Since our extension is intended to be used by users in their ecosystem, we need to make sure that we offer values in information in a non-intrusive way and minimize potential pitfalls.

Pull vs. push mechanism We think that it is important to lower the usage barrier as much as possible. Therefore, we use the push mechanism: users do not need to go to another site or make any queries to see annotations. This consideration leads us to develop a browser extension.

Since we do not know to what extent to which users care about the annotations, this issue leads us to the second important design consideration, which is clutterness.

Clutterness We do not want to pollute the page when users do not want to see too much extra information. Therefore, we use tooltips to allow users to get more information on

demand. Currently we highlight in yellow the texts that have annotations. In future iterations, we consider adding a little icon so that it is less visually distracting than yellow highlights.

Consistency Currently we choose to make all tooltips have the same format. Specifically, each tooltip will consist of one (possibly interactive) visualization, a title, and an explanation. In the future, we could consider allowing multiple interpretations, so we might allow each tooltip to be a list of cards. User testing can also help finalize the design.

Privacy and user messaging It is important to protect users' privacy. We need to be careful about how we use and collect personal information. Two important questions are the following: which information do we collect from users, and for what purposes? Collecting more information from users can allow us to do interesting analytics, but users might also be less willing to share information. We need to find the right balance. Currently, we do not collect any information into our system, which might not necessarily be the case in a later phrase. Regardless, it is important that user messaging is clear, so that users and developers both have clear understanding about risks and opportunities arising from using this tool.

Interaction with the original web page As is the case for any browser extensions, we need to make sure that our extension does not significantly slow down the page, or cause any errors. For CSS classes, we make sure that class names are prefixed by `pinfo_viz_` to avoid any conflicts.

RESULTS AND DISCUSSION

See Figure 3. In the 180 miles example, the following are reasons why the Yosemite tooltip annotation works well for me personally. First, I have been to Yosemite, so I can make sense of the distance from me (at Stanford) to Yosemite. Second, I do not have to go to another site to get annotations. Finally, if there were no well-known remarks at that distance, I could still see a circle with a 180 miles radius around me.

Current PinfoViz features

- Unit conversions between kilometers and miles are fully functional
- The regular expression pattern matching went through many iterations and seems to capture exactly the distance measures that we want. The extension highlights the distance text in yellow.
- On the test page and other simple pages, when users hover the yellow highlight, we show an interactive Google Map tooltip with a radius around user location.
- The map places a marker at the current user location.
- We also show the explanation relating to the Yosemite if the distance is in the [0.8, 1.2] times the distance from Stanford to Yosemite; otherwise, we do not show explanations.
- Users can select a location after clicking the browser icon.
- On the test page, users can log-in via Facebook to the PinfoViz facebook app, which allows us to get their basic personal information such as their public profile.

Current PinfoViz limitations

- Currently there is a bug, which is that the second Google Map that is being hovered does not render properly (showing a partial grey area). Refresh the page and hover over that one first to see proper rendering.
- Other pages should in principle work too, but currently there is a javascript dependency problem for some pages.
- Landmark energy function has not been implemented yet. This would require more work in getting the knowledge base, the baseline energy function, and all personal information ready in the right format. These are future work.
- Currently the browser extension can send a message to the content script, which allows personal information to change annotations. Currently we hard-code the location as Stanford and do not change location even though the settings have changed (due to the time constraint).
- Further Facebook permissions (especially user tagged places) are in review to get approved.

Discussion

To generate useful annotations, it is important that the algorithm understands the semantics of the text (see Future work). Semantic parsing in natural language processing is still far from human-level understanding, so this issue presents a great challenge. However, we could work around by presenting multiple options for interpreting.

There is an interesting consequence of the choice of the function f_u described earlier. If users only visit a few places, those will have very high scores, and therefore are much more likely to show up than other landmarks near them. The limited diversity might become an issue when users use the extension many times. There are two possible fixes: 1) lower the coefficient in front of n in the equation of f_u, and 2) introduce randomness into the system. Small stochasticity could lead to interesting but not-too-far-off results. The randomness here might also help users get out of their comfort zone a little and learn more about new landmarks.

USER TESTING

We interviewed around 5 users (friends and colleagues) and had over 40 students and guests from the Spring 2016 CS448b (Data Visualization class at Stanford) progress presentation and from the final poster session give feedback about the framework and the demo. The main focus in the interview was qualitative feedback: do you find this tool useful, and does the current design serve the purpose? The format of the interviews was a mix between our guidance and their independent exploration of the tool. Below is a summary that represents various factors that users think about when using this tool:

- [**User interface**] Most users think that this will be a useful tool for them. Many people especially liked the Chrome

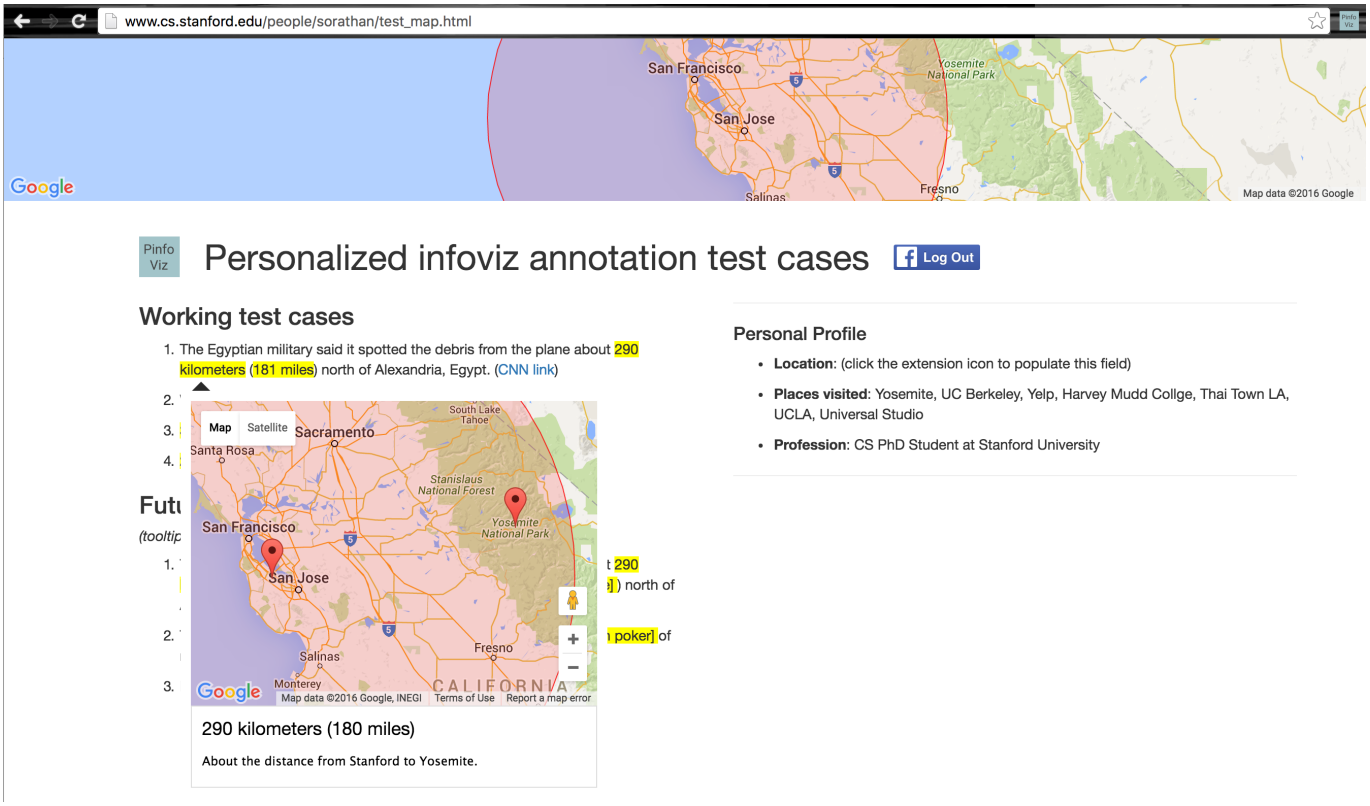

Figure 3. Example of the extension annotation in action. Features include unit conversion, an interactive Google Map as a tooltip, a spatial analogy description.

extension idea, as it is a "*good non-invasive way to present this tool.*"

A dual of opportunities is risks, which we will address now.

- [**User experience**] "*This is so cool! My only concern I have is that if I'm reading a site that has a lot of numbers that I might not actually care about knowing the length of, the article might get cluttered. Perhaps add a way to make sure the explanations of numbers doesn't get too cluttered.*"

  We note this concern and have addressed this issue in our design consideration.

- [**Personalization**] "*I really like the idea of using visualization to make unit conversions more understandable. I think temperature conversion would be super useful, perhaps more about the "feeling-temperature" not the actual (ex. It's 66 F degrees but it feels more like 60 F degrees which is chillier).*"

  This is an interesting comment, as it gives us an idea of another way that an annotation could be helpful: simplification. It is likely that "feeling-temperature" is dependent on users, so we think that our personalization framework is on the right track.

- [**Semantics and privacy**] "*I am curious about the decision to use facebook. It seems you are applying it as a proxy for the amount of information that a person can relate to. I would almost say that the places I grew up in, or read about in popular culture are just as formative as the places I have traveled to or posted about. Perhaps digging a layer deeper than facebook would give better personalizations.*"

  This is a surprising comment. We thought that getting information from Facebook has almost crossed the privacy line, but for this user it seems he or she is willing to share more information than Facebook.

- [**Value added**] A lot of people expressed that they would find this tool useful, but we would also like to get a sense of the extent to which users need this tool. To this end, we asked: "*If this tool isn't free, do you consider buying it?*" The response was a no, there wasn't enough value added for this particular user. But then after more conversation, we found out that she misunderstood the tool and thought she needed to go to a particular site to see tooltips working. She suggests that target users who would find this tool indispensable are advertisers, who need to make a living out of having people compare different numbers.

FUTURE WORK

Conduct more user testing

Previous user testing focuses on qualitative feedback. We intend to gather more qualitative evaluations as well as quantitative evaluations. One measure is the helpfulness rating

on the extent to which the annotations help users understand the numbers, and we can compare the rating with [5]. While self-report numbers are helpful, additional objective measures can help make the evaluation stronger. We plan to do user testing to test whether they can reconstruct the numbers more accurately when we present them with annotations.

Improve the current design

Due to time constraints, we have not implemented all features we think are helpful. For example, the annotation below is helpful because it also shows a route from Stanford to Yosemite, and also a driving time.

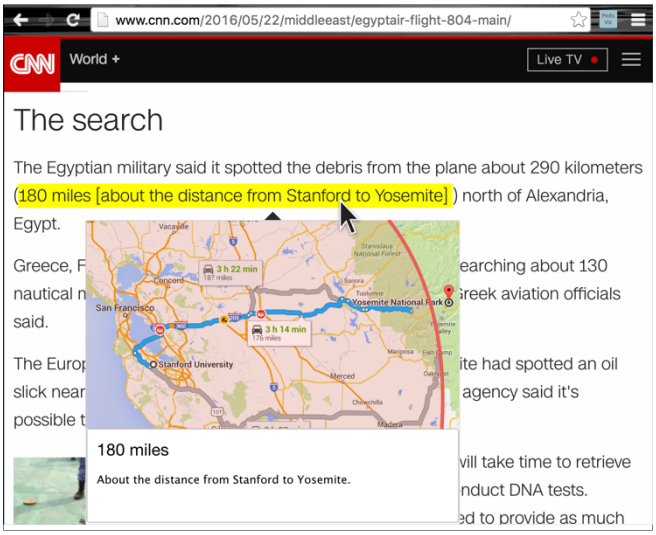

Figure 4. An improvement idea. Everything is live, except that the tooltip picture is a static picture drawn beforehand.

Further leverage personal information

An interesting question is how much personal information are users willing to share? We can design the extension in such a way that personal information only stays on the client side. That way, users do not lose their privacy. But this approach presents a challenge for us in making the extension self-contained.

Allow browser extension users to give in-site feedback

We could potentially allow users to rate our annotations. Together with their personal profile, their ratings can be illuminating. We can cluster users into different persona, and we can learn from this extension what kind of information each persona will find helpful.

Two considerations are a) this will need to be as frictionless as possible in order for users to enter data and for us to get enough signals, and b) the baseline annotations should already be somewhat useful for users to use the extension at all.

Take context surrounding numbers into account

For example, the word "Cambridge" itself might mean a river, a town in Massachusetts, or a town in English, among others. Even numeric measures such as "198 cm" could have totally different impressions for readers. For example, if this number represents a height of a person, then the person is extremely tall. If the number represents the height of a door, then it is very common.

Contextualize numbers by comparing with population statistics

Other examples include contextualizing the number of application downloads by comparing with the top downloaded applications. Compare prices with items that users usually buy or likely know prices.

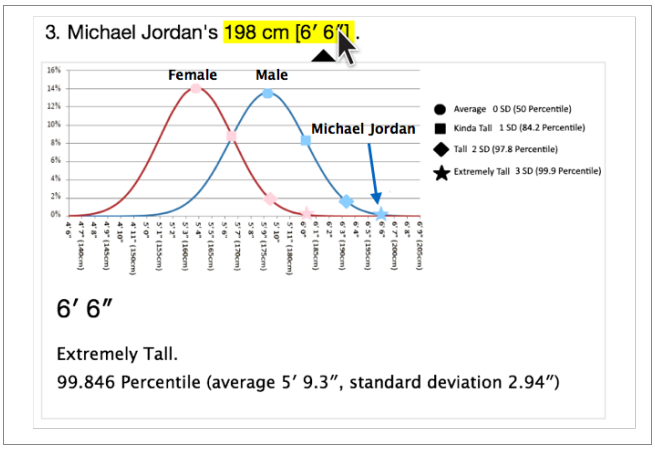

Figure 5. Put height in context.

Figure 5 describes the height of the basketball player Michael Jordan in the context of the population height distribution for males and females. The plot indicates where Michael Jordan is in the distribution, and the text explanations include a percentile number and a simple interpretation ("Extremely tall"). User testing session reveals that users also want compare Michael Jordan's height with the height distribution of NBA players, and also the user height.

Expand to more domains (guided by research and user testing)

We currently discuss spatial analogies as our main task in this paper, but personal information can be used to help annotate numbers in a much wider context.

One simple extension is to expand to other physical quantities, such as ratios. Imagine the following situation: a Stanford plant physiology PhD student studies Atomic Force Microscopy, and in his thesis defense, he needs to describe that the needle in the tool he is working with is about 1000 times the size of the study subject, so he has to be extra gentle not to break the target. At first, the audience gets an idea that this ratio is quite a big number, but it is still vague how big it is. Then he goes on, "*that's equivalent to 3 Hoover towers over a basketball*." The entire room laugh, but also have a vivid picture of this ratio. Note that being in Palo Alto is not

sufficient to ensure that the audience know about the Hoover tower, but being a Stanford student is. Imagine being able to scale this kind of personalized analogies across a wide range of quantities. (Note: the scenario in this paragraph was actually a real story, when I attended my friend's thesis defense; credit: Witchukorn Champ Phuthong).

We also plan to look at the statistics of mentions in major online news sites and see which ones are the most popular ones. Among those, we will conduct user testing to determine which numbers are confusing. For example, we sense that the most frequently mentioned numbers in news are about people (e.g. the number of deaths, the cost of damages). The number of deaths can be put into perspective by relating to the number of deaths in other major events. The cost of damage can also be done similarly.

Empower writers to write more clearly or to personalize writing

This tool can be useful not only to readers, but also to writers. Annotations can help the writer concretely see alternatives, and can allow them to improve the clarity of writing overall.

Writers who want to write to, say, a Chinese audience may enter a Chinese persona into the Chrome extension and see which annotations are helpful to this audience. Then the writer can personalize this piece of writing to the Chinese readers without needing them to install this extension.

ACKNOWLEDGMENTS

Thanks to Maneesh Agrawala for helpful discussions and ideas, and to the Spring 2016 Data Visualization class at Stanford (CS448b) staff and students for helpful feedback.

REFERENCES

1. 2005. Google Maps APIs. https://developers.google.com/maps/. (8 February 2005). Accessed: 2016-06-05.
2. 2009. Wolfram|Alpha. https://www.wolframalpha.com/. (15 May 2009). Accessed: 2016-06-05.
3. 2012. TippedJS. http://www.tippedjs.com/. (1 January 2012). Accessed: 2016-06-05.
4. A. Chaganty and P. Liang. 2016. How Much is 131 Million Dollars? Putting Numbers in Perspective with Compositional Descriptions. In *Association for Computational Linguistics (ACL)*.
5. Yea-Seul Kim, Jessica Hullman, and Maneesh Agrawala. 2016. Generating Personalized Spatial Analogies for Distances and Areas. In *ACM Human Factors in Computing Systems (CHI)*. http://idl.cs.washington.edu/papers/spatial-analogies