A Millennium of Battles

Marguerite Delcourt
marguerite.delcourt@epfl.ch

David Froelicher
david.froelicher@epfl.ch

Christian Mouchet
christian.mouchet@epfl.ch

Abstract

For centuries, battles have been and still are the most direct action taken by established powers to project or extend their influence. By collecting, processing and analyzing data from the open source, collaborative Wikipedia platform, we are able to better understand how battles were fought throughout the last millennium. In this document, we report our findings from an analysis made over more than 7000 battles. We expose trends in duration and casualties over the years. Additionally, we show that the latter is the most important battle's feature in determining the short-term winner which is not necessarily the long-term one.

1 Introduction

In 1943, the soviets won the Battle of Stalingrad against the nazis while loosing 800,000 soldiers, twice the number of german casualties. In spite of these heavy losses, this soviet victory is considered as a major strategic turning point that lead to the soviet victory against the nazis' invasion. Human history contains numerous examples such as this one that show the importance and complexity of battles.

While battle-related casualties are far from being representative of all the damages caused by war, they have the particularity of being the result of a strategic and tactical planning, and are therefore worth analyzing. In this work, we make a first step towards understanding the evolution of the way powers wage battle by studying the trends and relationships between multiple battle-related features such as the initial strengths, the number of casualties or the result type of a battle.

In order to support these observations, we first describe how we collected and processed data on more than 7,000 battles from the English Wikipedia database dump[1] and provide a short spatial and temporal descriptive analysis of the resulting dataset. Then, we report on our findings by observing trends in duration, number of casualties and how indecisive the battles tends to become over the last thousand years. We isolate the adversary losses as a critical feature determining probability of victory and show that this is interestingly not the case for longer-term victories. Lastly, we compute cumulative time spent in battle for each of our extracted belligerents and provide a ranking of the most warlike ones.

2 Related works

The Uppsala University has a Conflict Data Program[2] which is the oldest and principal provider of war related data. It is considered as the global reference in terms of armed conflicts studies. Even though this ongoing project operates at a much larger scale, their results correlate with ours (mostly for peak values). The main difference lies in a finer granularity of considering ”violent events”, which include but are not limited to battles, resulting in differences in trends.

Another interesting related work is the ”War and Peace” article from the *ourworldindata.com* website[3]. Their conclusions agree with ours to state that the past was not more peaceful than the contemporary era. Their data also show that the common belief that we live in the most dangerous times is the mere consequence of our human tendency to give less importance to past conflicts than present times ones.

3 Data collection and processing

Given the semi-structured nature of our base dataset, data collection was a significant part of our work. The task was to extract clean and normalized features for each battle related page in the English Wikipedia dump file (∼44 GB). Based on the assumption that the data collection would certainly be an iterative process, we split our pipeline in three steps, enabling each of them to be ran separately and avoiding larger scale computation to be required at each iteration.

[1] https://meta.wikimedia.org/wiki/Data_dumps

[2] http://ucdp.uu.se/

[3] https://ourworldindata.org/war-and-peace/

3.1 Page extraction

The first pipeline step is the only one to be run on the cluster. It performs the page selection based on a regular expression matching of the page title, and outputs a consolidated JSON file containing all selected pages. By avoiding further processing at this step, we effectively reduced our dependency on the cluster, the output being already small enough for local processing ($\sim$123 MB, 27k pages, which was way beyond our initial estimate of 10k).

3.2 Field extraction

We leverage on the presence of an *infobox* in most of the battle-related pages. This pipeline step looks for it in the page's source code, which is written using the Wikitext[4] markup language. The syntax generating the *infobox* template[5] provides a *key-value* set for each battle where the keys are conveniently standardized (e.g. *date*, *combatant1*, *combatant2*, ...). Values, however, are free text fields, usually containing Wikitext, and natural language formatted for humans. For this reason we qualified our dataset of being semi-structured. The output of this part is a file containing one JSON per battle, with its extracted *key-value* set, thus reducing the size to $\sim$14 MB. The number of pages where an *infobox* could be found was 7486, which is within our initial range estimate, and still comfortably large. It appeared that many pages were in fact redirects, drafts or discussion pages.

At this point, we were able to evaluate which fields could be turned into features by checking the number of concerned pages and the Wikitext processing feasibility.

3.3 Feature extraction

The last pipeline step consists in turning *key-value* pairs from the previous step into clean and normalized features. This was clearly one of the most time consuming part of this project, as each field contained very different types of data and thus had to be considered separately.

date is a field of primary importance because we initially proposed to analyze the duration of battles. The main difficulty was to deal with many representations of date ranges, as well as dates happening before year 0, which python does not easily accommodate. We decided to ignore such dates after noticing they represent a very tiny fraction of all battles.

combatant_*n*: are the fields containing name(s) of the involved belligerents. Most of the battles have them populated for $n \in \{1,2\}$, representing the two sides. The difficulty here is twofold: first, the feature type needs to accommodate for coalitions of multiple parties. Second, their affiliation has to be captured at the right level of granularity. We find that a simple yet precise way of learning the state affiliation is to use the flag icons displayed next to the combatant name.

result's content appears to vary a lot from battle to battle, ranging from strait naming of the winner to elaborated aftermath analysis. Consequently, we choose to focus on extracting the victory *type* as it fits well in a categorical variable. In addition, we attribute the extracted result to one of the combatants based on string similarity. This give satisfactory results in most cases, enabling meaningful analysis within sufficiently large sets of battles.

coordinates fields are quite rarely populated, which caused problem when studying the spatial distribution of battles. We overcome this issue by extracting page-level geo-tagging data in the previous pipeline step.

strength_*n* contains a summary of side n's strengths, usually in number of soldiers, but often including weapons and assets such as tanks and ships. In most cases, we are able to extract the soldier count or an estimate of it (in case where ranges of values are provided).

casualties_*n* is the most difficult field to extract normalized information from. Even though it almost always contains a casualties amount breakdown between wounded, killed and captured, the semantical meaning of this breakdown appears very hard to capture. For example, many pages provide breakdowns coming from multiple conflicting sources, when others exhibit phase (or day) based casualties count. On the other hand, conjunction and disjunction used to attribute numbers to the casualty type are often difficult to make sense of. We decide to focus our analysis on one single *casualties_n* feature per combatant, obtained by summing all values (killed, wounded, missing, captured, disappeared).

4 Descriptive analysis

While providing exhaustive descriptive analysis for every features is not feasible within this short report, given their diversity, the reader can refer to the companion notebook. In this section, we focus on analyzing the spatio-temporal coverage of our dataset, as it is important for the rest of the discussion.

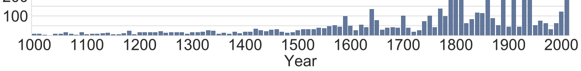
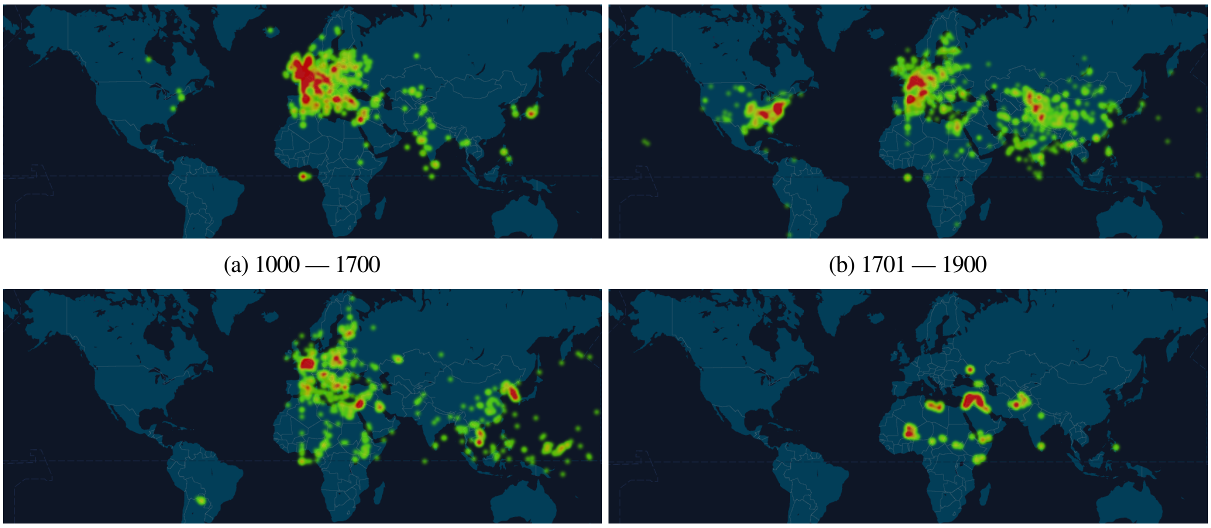


Figure 1: Battle count per decade

We first look at the distribution of events in time in Figure 1. Peaks are clearly observable during the

[4] https://www.mediawiki.org/wiki/Wikitext

[5] https://www.mediawiki.org/wiki/Help:Templates/fr



(a) 1000 — 1700

(b) 1701 — 1900

(c) 1901 — 2000

(d) 2001 — 2017

Figure 2: Geographic density of battles for different year ranges.

American Civil War, the first and the second World Wars, while the years 1000 to 1700 have low battle counts per decades. Notice that this does not mean they were more peaceful, only that less battles were recorded.

Then, we display the geographic distribution for chosen periods and as we can see in figure 2, battles were occurring mostly in Europe between the 11^{th} and the 17^{th} centuries. The 18^{th} and the 19^{th} centuries are the only ones in which many battles occurred on the North American continent (with the 1861 Civil War), these centuries also witnessed the First Opium and Sino-Japanese Wars in Asia.

5 Results

In this Section, we report on our analysis of the dataset we built, focusing on the most interesting findings. The interested reader will find our data exploration and extended analysis in the companion notebook.

5.1 Duration of the Battles

We observe that the duration of the battles over the last millennium increased almost monotonically. In fact, we can see in Figure 3 that the average duration of a battle has never been higher than nowadays. We show here the duration in order of magnitude. This is because the distribution of the duration values was concentrated on small values with a heavy tail going to extreme values. Thus, using the average of the logarithm duration values, we attenuate the influence of extreme values while showing an augmentation of almost one order of magnitude in the battles' duration. Notice that the average duration of a battle was almost 34 days during the XX^{th} century while it is currently of 64 days in the XXI^{st} century.

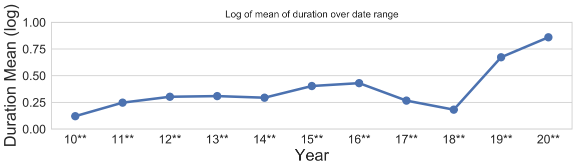


Figure 3: Mean of the duration of battles in the last thousand years by century on a logarithmic scale.

5.2 Evolution of the Casualties

In Figure 4, we notice that the percentage of soldiers engaged in a battle that are wounded, killed, captured or that disappeared decreases throughout the years. We observe that, following the same trend as shown in Figure 3 for the battle's duration, the percentage of casualties decreased before increasing in the XX^{th} century. We can infer that in both cases this is due to the world wars because they contain numerous battles which made countless casualties, notably because of technology advances such as aerial forces or gaz attacks.

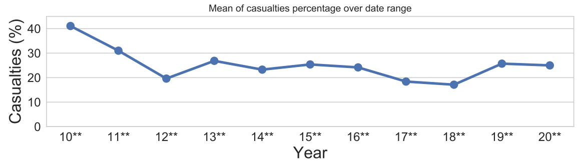


Figure 4: Mean of the percentage of casualties in the soldiers.

5.3 Indecisiveness of the Battles

We observe, in Figure 5, that battles tend to be more indecisive nowadays than they were in the past. In fact, the fraction of indecisive battle within one century is increasing slowly since year 1200 and at a much higher rate over the last two centuries. This can be explained by two factors: the first being that battles in the further past were most likely reported by the winner with probably less accuracy and objectiveness than they are now. The second explanation lies in operational theaters (like cities) and tactics enabled by modern weaponry such as long range weapons. Ultimately, longer distances between the combatants and modern combat vehicles greatly enhanced the ability to retreat.

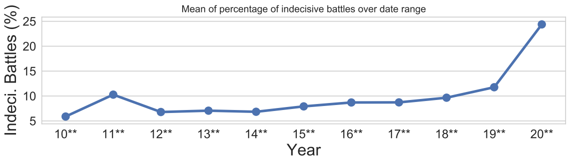


Figure 5: Mean of the percentage of indecisive battles over the last thousand year.

From Figures 3, 4 and 5, we conclude that battles are longer, make less casualties and are more indecisive over the years. This supports the fact that nowadays the battle's purpose is not only to invade another territory by beating the other combatant. In fact, modern battles seem to be more complex in the sense that they often serve a higher strategic or even political goal and have their result studied in many different axis.

5.4 Outcome of the battles

We continue our analysis by studying the advantage provided to a combatant that has fewer of casualties, larger strengths and lower casualties to strength ratio. We compute the percentage of battles won by a combatant that had at least one of these advantages, grouped by three different victory types: *tactical*, *strategic*, *decisive*. The *tactical* level is the lower one in the military level of planning[6]. Therefore, involving narrower scope decisions and shorter-term consequences. On the other end, the *strategic* level is the highest one, involving longer-term consequences.

We first observe that an advantage in strength only is not sufficient to increase one's chances to win a battle, as this would completely disregard other tactical advantages of the opponent. On the other hand, the a-posteriori advantage in the casualties (both in ratio and absolute value) is much more significant, as it is present in around 75% of the victories for all but strategic ones. Interestingly, we observe that for this victory type, an advantage in absolute casualties has much less effect on the outcome (in fact, it looks like it is the opposite), and a little less effect in the ratio advantage. This would typically generalize the mentioned case of the Battle for Stalingrad, where the Nazis' strategy was not able to provide victory even though it was tactically superior. Figure 6 illustrates these observations.

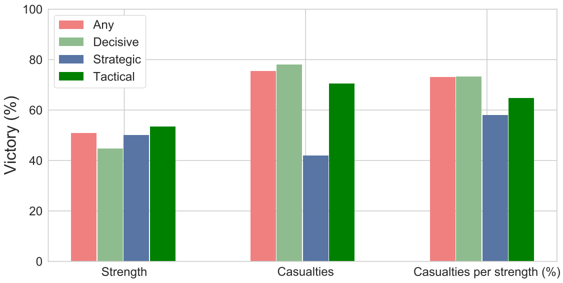
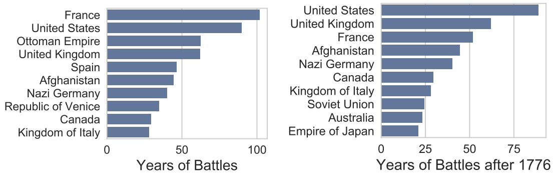


Figure 6: Victory percentage with advantage in either strength, casualties or casualty ratio.

5.5 Years of Battles per Countries

In Figure 7, we observe that among all countries in our dataset, France is the one that fought the most. Nevertheless, the United States, which are a major actor in the international history of battles, were only created in 1776 when they proclaimed their independence. Thus, it is interesting to do the same ranking starting from this year.



(a) 1000 — 1700 (b) 1701 — 1900

Figure 7: Cumulated duration of battles engagement per country. From 100 in (a) and from 1776 and the United States independence in (b).

These results tend to support the common belief that the United States are always in war. In fact, as we observe in Figure 8, they were engaged in a at least one battle per year for more than 150 years in 242 years of existence.

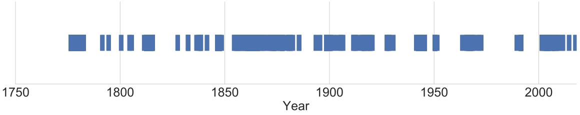


Figure 8: Timeline of the USA engagement in battles.

6 Conclusion

In this report, we have collected and processed the data from Wikipedia in order to compile a comprehensive and usable dataset on battles throughout the last millennium. We have shown that battles have become lengthier and more complex, with the probably increasing difficulty of achieving tactical and strategic goals resulting in more indecisiveness. We also pointed out that, while the number of battle-related casualties decreased, the casualties-to-strengths ratio remained almost constants, and that these two statistics are critical for the outcome of the battle, even though there are not enough to achieve a major strategic victory.

[6] http://www.esquire.com/news-politics/politics/news/a39985/four-levels-of-war/