Collective Memory in Poland: a Reflection in Street Names*

Radoslaw Nielek[1], Aleksander Wawer[2], and Adam Wierzbicki[1]

[1] Polish-Japanese Institue of Information Technology,
ul. Koszykowa 86., 02-008 Warsaw, Poland
[nielek,adamw]@pjwstk.edu.pl
[2] Institue of Computer Science Polish Academy of Science
ul. Jana Kazimierza 5, Warsaw, Poland
axw@ipipan.waw.pl

Abstract. Our article starts with an observation that street names fall into two general types: generic and historically inspired. We analyse street names distributions (of the second type) as a window to nation-level collective memory in Poland. The process of selecting street names is determined socially, as the selections reflect the symbols considered important to the nation-level society, but has strong historical motivations and determinants. In the article, we seek for these relationships in the available data sources. We use Wikipedia articles to match street names with their textual descriptions and assign them to the time points. We then apply selected text mining and statistical techniques to reach quantitative conclusions. We also present a case study: the geographical distribution of two particular street names in Poland to demonstrate the binding between history and political orientation of regions.

Keywords: collective memory, Wikipedia, street names

1 Introduction

The idea behind this article is based on an observation that the choice of street names is a reflection of selective properties of national collective memory. The choice of street names reflects what is worth remembering and what is important, how we want to remember the past.

The notion of collective memory comes from a French sociologist Maurice Halbwachs [5]. He distinguishes three types of memory: autobiographical (personal, individual memory), collective (group memory that maintains society's interpretations of the past) and historical (shaped by historians).

Collective memory has been the subject of numerous scholarly publications in the fields of sociology, cultural anthropology and history, too numerous to list here. The attempts to investigate it using computational means are relatively new and few, mostly associated with text mining historical corpora.

* This work is supported by Polish National Science Centre grant 2012/05/B/ST6/03364

The studies of street names distributions are exclusively the domain of computational cartography. For instance, [8] analyze patterns and relations underlying the selection of European cities as names of German streets. Their analysis is focused on spatial proximity and does not consider historical factors (as the authors explicitly put it: *factors like bilateral or historical relations of cities are relevant for selective cities and can be seen as noise*). Takahashi et al. [7] has tried to predict importance of historical events based link structure of Wikipedia entities. Application of NLP tools and text mining approach for studying historical documents has been quite well researched starting from five thousand years old Sumerian clay tablets[6] through 19th and 20th century books [3] and ending with computational history of the ACL [1]. Au Yeung et al. [2] have tried to study collective memory about historical events by extraction references to the past in news articles.

The paper is organized as follows: Section 2 describes the data sources used in the analyses, Section 3 describes historical mappings of streets names. Section 4 describes the result of machine learning and text mining experiments on Wikipedia articles. Section 5 is the case study of geographical distribution of two historically-linked street names. We conclude in Section 6.

2 Dataset

Dataset with names of all streets in Poland has been obtained from the TERYT[3] as a XML file. Every street is accompanied with three numbers that identify province, district and community. Number of all streets exceeds 247 thousands but only ca. 35 thousands names are unique (the most popular street name is Polna[4] and occurs 3132 times). Words like street, square, boulevard etc. have been removed from the dataset because of two reasons. First, differences between street, alley or square are not important for our research goal (we do not want to treat separately same street patrons because in some cities their name is given to street and in other to square). Second, street/square/boulevard prefix is useless for matching street names with Wikipedia entities. All street names composed of less than two words were removed to clear the dataset from ordinary words like Green Street or Long Street (and leaving only historical events and names; in Polish, combinations of name and surname are used to identify people, streets and all historical events are also named with the use of at least two words). After the data clearance the dataset contained 16635 street names matched with entries in the Polish Wikipedia.

The process of matching Wikipedia entries with streets names faced many surprising difficulties. First of all, street names in Polish are in other casing than the titles of entries in Wikipedia. Therefore, an automatic conjunction of all street names was needed (it is not a trivial task in Polish as many nouns and names are irregular and many people's names have foregin roots). The dataset

[3] TERYT is a National Official Register of the Territorial Division of the Country and is used by Polish administration for many location-related services.

[4] The adjective polna can be translated as field or wild.

contains also some misspellings but even bigger issue was an omissions of second given names and quite random order of names and surnames. Therefore, we ignored word order. Yet another problem was caused by abbreviations in names and titles.

To address previously mentioned issue,s a special function that measure similarity between street name and entries has been designed. Streets have been connected with the most similar entity, but only if similarity measure exceeded certain threshold (selected experimentally). Additionally to automatic matching, all non-matched streets and 25% of matched streets (selected according to a frequency of occurrence) have been checked manually. The final dataset[5] contains 8060 street names with their corresponding entries in Wikipedia. For the remaining 8575 street names either no corresponding article on Polish Wikipedia existed or matching could not be done unambiguously.

The drawback of the approach presented in previous paragraphs is that results may be biased by an existence of entries on the Polish Wikipedia (some type of people or topics may be systematically omitted by Wikipedia editors). Wikipedia is known as high quality (but not error-free), extensive encyclopedia and the Polish version is one of the biggest over the world with over one million of articles. Additionally, as neither sentiment nor opinion are extracted and used in this research even biased, bad quality articles with a lot of omissions may be useful. On the other hand a manual check of the matching results has revealed that for some street names corresponding article does not exist. A closer look on biographies of these missing people may be extremely interesting task but has not been done as a part of this study.

3 Street Names and Historical Links

From each Wikipedia article associated with a street name we extracted a set of year-alike numbers using regular expressions. We sorted the list and computed its median value as the most representative for an article. The experiments demonstrate that the method is nearly error-free in matching street names to historical periods and selecting the most prominent dates.

Figure 1 presents street frequencies plotted against the time line of extracted median years.

Dates related to street names cannot be seen as an indicator of when particular street has been built. Names for streets are selected from the whole Polish history. The only rule is that people are honored with naming streets only after dead.

Generally, it appears that the number of street names related to dates before 1800 is very low. Each street name, mapped to the period before 1500, represents either a king or a queen. The period of 1500-1700 is marked by two writers (Rej and Kochanowski), one artist (Stwosz) and two kings. Interestingly, only successful, militarily victorious rulers appear on the list.

[5] The final dataset is publicly available and can be downloaded from: url-http://nielek.pl/histoinformatics2013/8060streetnames.csv

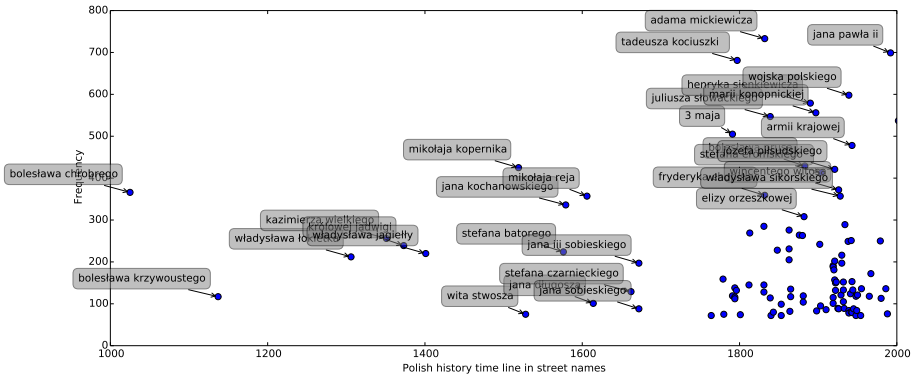

Fig. 1. Frequencies of street names mapped on historical time line. To improve readability only the most common name in given periods were labeled.

The period after 1800 is entirely different: there are many more associated street names and their structure is distinct. The period of 19th century is often considered as the time of national awakening, the raise of social programs to establish unique national identities, connected to language and rooted in local symbolic space (nation-language states). Therefore, it is hardly surprising that the most frequent street names (matched to the period after 1800) belong to the spiritual leaders such as romantic poets (Mickiewicz and Slowacki) and the pope (John Paul II of Polish origin), the most notable romantic composer (Chopin) and a novel writer (Sienkiewicz). All these figures and their works form a symbolic space (a canon) which defines Polish nationality and is a vehicle of national identity. It forms what might be considered a body of romantic nationalism, a developing imagined community of Poles with its spiritual value, unique character and language.

The timing of such ideas is not specific to Poland and shares many similarities with other countries. In fact, many nations were created using principles based on romantic nationalism as the source of legitimacy, emphasizing the binding between language and nation.

The period after 1800 overlaps with important changes in social and economic structure: industrialization. The intensification of nation-building efforts and rise of nationalism in this specific period are, according to Ernst Gellner [4], a side effect of industrial changes. In order to ensure availability of technically skilled work force, a state must enforce cultural standardisation, which is achieved by unification of political and the national (nationalism). The frequency of street names associated with nation-building leaders is therefore a consequence, or a reflection of pushing to unify the national and political, and as such fits in Gellner's conceptual framework.

4 Mining Street Name Frequencies

This section describes the result of machine learning and text mining experiments on the Wikipedia articles.

4.1 Predicting Street Name Frequencies

Number of street name occurrence may be seen as a rough estimation of importance of given person or event. As particular street name can appear only once in each city (there are some exception when one person appears under two names, e.g. Karol Wojtyla and pope John Paul II), many occurrences of the same street names in the dataset require that a lot of local communities decide to honor this particular person or event. The aim of the research presented in this section is to check whether the importance of people and historical events for society can be predicted based on features extracted from the Wikipedia.

In the experiment, we use 8060 unique street names with corresponding Wikipedia articles. We focus on predicting how many times does a street name occur knowing only article text and variables computed over the text. The problem may be seen as a regression and thus appropriate performance measurement metrics include R^2 and Mean Squared Error (MSE).

We computed following feature spaces:

- F1: counts of positive sentiment sentiment (npos) words, negative sentiment (nneg) words and total number of words (ntot);
- F2: number of extracted years (nyrs), aggregated negative (sneg) and positive (spos) word sentiment, finally total number of words (ntot);
- F3: lexemes (unigrams) as bag-of-words (disregarding word order) vectors, TF-IDF weighted, computed on word base forms obtained by the means of morphological analysis.
- F4: wikipedia categories for each article.

Table 1 presents the results obtained using three regression algorithms: ordinary least squares linear regression (Linear), Ridge Regression (Ridge) and Support Vector Regression (SVR) – with a radial basis kernel of degree 3. All results were computed as average values in 4-fold cross-validation.

Algorithm	Linear		Ridge		SVR	
Measure	R^2	MSE	R^2	MSE	R^2	MSE
F1: npos, nneg, ntot	0.0256	910.00	0.0288	910.41	0.0202	950.24
F2: nyrs, spos, sneg, ntot	0.0288	910.41	0.0288	910.41	0.0206	950.24
F3: all lexemes	n/a	n/a	0.1052	1000.48	0.0260	954.99
F4: wikipedia categories	0.001	933.19	0.001	933.19	0.0246	954.99

Table 1. Numbers of words and annotated text fragments.

The results demonstrate that on dense feature sets (F1 and F2) simpler algorithms such as least squares linear regression or ridge regression perform better. The improvement is apparent in both MSE and R^2. The SVR algorithm outperforms two other regression types on sparse and large data of lexemes (F3). However, the results are very preliminary and need a lot of fine-tuning, perhaps backed by other data sources, to achieve satisfactory performance. The values of R^2 are generally low. On average, the predictions are mistaken by around 30 street name occurrences (MSE).

4.2 Predicting Street Names from Wikipedia Texts

In the second experiment, we focus on predicting whether a Wikipedia article has at least one associated street name (regardless of actual frequency, provided it is non-zero). We begin with the same set of 8060 Wikipedia articles with corresponding street names, but narrow it to a subset of 5698 biographies (we remove all articles that are not biographies). We also pick a random set of the same size of articles – biographies linked to history of Poland, but without a corresponding street name.

The question here is a similar one to that of the previous section: is textual information of Wikipedia articles sufficient to predict whether there is a street name linked to the article. This time, the problem may be seen as a classification and thus appropriate performance measurement metrics include precision and recall.

We use feature space of all lexemes, called F3 in the previous section, and a Logistic Regression classifier. We report the results as averages in 10-fold cross-validation in Table 2.

Measure	Precision	Recall
Result	0.841	0.875

Table 2. Precision and recall of predicting street names from Wikipedia biographies.

As much as the results of regression on street name frequencies appear preliminary, the classification experiment reported in this section seems very promising, as the baseline of a balanced data set (as in this case) is at 0.5. Wikipedia articles carry relevant information and the models predict the existence of street names with good performance.

5 1st vs 3th May

1st May (also known as "May Day") is ancient spring festival in Northern Hemisphere but also the International Workers Day. May 1st was established in 1950 as a public holiday in Poland and used to be the most important public holiday in

communist-era in Poland. The celebration usually took the form of huge streets parades (hundreds of thousands participants). Participation in such parades was often seen as a support for communist government. Although it is not officially celebrated any more, it is still a public holiday. Two days later in calendar there is another important day for Poland. 3th May 1791 "the first constitution of its type in Europe[6]" has been declared in Warsaw. From 1919 till 1951 (with a break during the Second World War when Poland was occupied by Germany and Russia) 3th May was a public holiday. Polish communist government after students unrests in 1946 has forbidden an official celebration of this holiday. In 1990, one year after the fall of communism in Poland, 3th May has regain its status as a public holiday.

Attitude toward these two holidays is sometimes seen as rough approximation of evaluation of communist-era in Poland. Street names that honor International Workers Day used to be very common in Polish cities between 1950 and 1989 (quite often it was one of main streets in cities). After the fall of communism, 1st May streets have started slowly to disappear and new street names honoring anti-communists activists (e.g. Marshal Jozef Pilsudzki) and ideas that were fight back by communists (e.g. declaration of 3th May constitution) are getting more popular.

The process of changing street names is quite slow because it has to overcome peoples habits and status quo dictate. Additionally, there are not many completely black or white people in Polish history and their evaluation vary strongly. Crucial factors in evaluating events in history are political views. In Poland, as in almost all countries[7], exist very stable geographical patterns of political support. Western provinces tend to support left wing parties and Southeastern provinces are more conservative. According to Polish law local councils decide about street names, so the geographical patterns of political support should be also visible in a ratio of communist to anti-communist names.

To verify the hypothesis stated in previous paragraph two street names have been selected 1st May (communist) and 3th May (anti-communist). The ratio of anti-communist vs. communist names for provinces varies really strong from 0.27 for Opolskie (southwestern part of Poland) to 2.58 and 3.9 for Podkarpackie and Maopolskie (southeastern) respectively. Spearman correlation between calculated ratio and support for left/right wings candidates in the second round of the last presidential election[8] is 0.504. On fig. 2 a characteristic spatial pattern is visible where western and southwestern provinces are more left-oriented than center and eastern provinces.

[6] The first modern constitution was declared in the USA. The well-known French constitution was declared three months after the Polish constitution.

[7] For example in the USA democrats never win in so called redneck states like Texas or Kentucky and in Germany CSU (right wing party) rules in Bayern over 50 years in row.

[8] The last presidential election in Poland took place in April 2010. In second round voters could choose between Jaroslaw Kaczynski (right-wing party) and Bronislaw Komorowski (central/left-wing party).

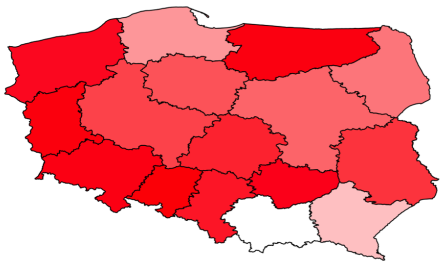

Fig. 2. Number of 3th May streets in comparison to 1st May streets in provinces. Darker color denotes lower value.

6 Conclusions

This paper describes computational experiments aimed at using a database of street names as a resource of investigating past, as reflected in collective memory of a nation.

The analyses described in this paper prove that street names are an important carrier of national identity and have strong historical connections. Using automated mapping of street names to their associated dates, we examine their distribution in time. We observe that the type of symbols, selected as street names, as well as their density, changes in time. The change overlaps with industrial revolution and may confirm Ernst Gellner's views on the birth of nationalism.

Dataset crafted for the research presented in the paper creates enormous opportunities for further studies. Connection of street names and demographic information about cities and communities may reveal interesting patterns. Exchanging an existing administrative division of Poland with it historical versions is another step worth considering. Furthermore, frequency of particular street names has been used as a rough estimation of street name importance but it may happened that some names are used very often but only for small streets on suburbs. Therefore, a metric that combines frequency and geospatial features (e.g. location, length) will be in future examined. Very interesting results can also be obtained by analyses whether selected street names are relating to local or national or global level. Using additional sources of historical information, next to Wikipedia, may improve coverage and add some missing features but will likely make an automatic analyses a nightmare.

References

1. Anderson, A., McFarland, D., Jurafsky, D.: Towards a computational history of the acl: 1980-2008. In: Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries. pp. 13–21. ACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012), http://dl.acm.org/citation.cfm?id=2390507.2390510

2. Au Yeung, C.m., Jatowt, A.: Studying how the past is remembered: towards computational history through large scale text mining. In: Proceedings of the 20th ACM international conference on Information and knowledge management. pp. 1231–1240. CIKM '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2063576.2063755
3. Gander, L., Lezuo, C., Unterweger, R.: Rule based document understanding of historical books using a hybrid fuzzy classification system. In: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. pp. 91–97. HIP '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2037342.2037358
4. Gellner, E.: Nations and Nationalism. Cornell University Press (1983)
5. Halbwachs, M.: La Mmoire collective. Presses Universitaires de France (1950)
6. Jaworski, W.: Contents modelling of neo-sumerian ur iii economic text corpus. In: Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1. pp. 369–376. COLING '08, Association for Computational Linguistics, Stroudsburg, PA, USA (2008), http://dl.acm.org/citation.cfm?id=1599081.1599128
7. Takahashi, Y., Ohshima, H., Yamamoto, M., Iwasaki, H., Oyama, S., Tanaka, K.: Evaluating significance of historical entities based on tempo-spatial impacts analysis using wikipedia link structure. In: Proceedings of the 22nd ACM conference on Hypertext and hypermedia. pp. 83–92. HT '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/1995966.1995980
8. Thiel, S., Pippig, K., Burghardt, D.: Analysis of street names regarding the designation of cities. In: Buchroithner, M.F. (ed.) Proceedings of the 26th International Cartographic Conference. International Cartographic Association (2013)