The Making of My Mother's Book: Named Entity Recognition for the Index of Persons

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences

BachoTEX 2024,
Bachotek 1–5.05.2024

Problem statement

Barbara Bielawska-Dębowska (1937–2020), my mother, did family research ($>$ 2300 persons in her data base), authored or edited six books about the family history:

1. Helena z Jaczynowskich Roth, Czasy. Miejsca. Ludzie. Wspomnienia z Kresów Wschodnich, Wydawnictwo Literackie, Kraków 2009.
2. BBD, Piotr Roth-Jaczynowski, O polskich Rothach, wyd. Piotr Roth-Jaczynowski, Łódź-Warszawa, marzec-lipiec 2000.
3. BBD, Wesołowscy z Rawy, Wydawnictwo WAM, Rawa Mazowiecka 2013.
4. BBD, Wesołowscy. Post scriptum, wyd. BBD, Łódź 2017.
5. BBD, Westernalia, czyli 10 pokoleń Westerskich, wyd. BBD, Łódź 2019.
6. BBD, Bielawscy – Pięciu w linii prostej, wyd. Łukasz Dębowski, Łódź, 2021.

Advantages of a brave decision

I converted my mother's MS Word file to LaTeX via HTML:

- High quality of typesetting for book publishing.
- Easy processing of hundreds of images.
- Automatically generated TOC.
- Automatic citations for a bibliography.
- Automatically generated family trees. (another talk!)
- Automatically generated index of persons. (really???)

Let's see what we need

```
Bielawscy, pięciu w~linii prostej:
\index{\textsc{Bielawski} Józef \textit{Józefat}}Józef --
\index{\textsc{Bielawski} Feliks \textit{Felix}}Feliks --
\index{\textsc{Bielawski} Ksawery}Ksawery --
\index{\textsc{Bielawski} Marian}Marian --
\index{\textsc{Bielawski} Jerzy \textit{Jurek}}Jerzy.
Ich życiorysy wypełniają ponad półtora wieku od 1772 roku
do 1940 roku.
```

How to generate so elaborate index labels automatically?

Named entity recognition

It may seem as simple as the following:

1. Get a list of given names and surnames.
2. Seek for patterns "(given name)+ (surname)?".
3. Put an index label for each mention of such a pattern.

Challenges

Normalization: "Jerzy Bielawski" can be referred to as:

- Jerzy, Jerzego, Jerzemu, Jerzym.
- Jurek, Jurka, Jurkowi, Jurkiem, Jurku.
- Jerzy Bielawski, Jerzego Bielawskiego, etc.

Disambiguation: "Helena", "Heleny", "Helenie", etc. refer to:

- Helena Bielawska-Barcikowska.
- Helena Hurysz.
- Helena Jaczynowska-Roth.
- Helena Lubicz-Płodowska-Westerska.

Idiosyncrasies of diminutives/pseudonyms/spelling:

- "Anoda" is "Jan Rodowicz".
- "Halina" can be "Helena Bielawska-Barcikowska".
- Gabryela, Xawery, Jakób = Gabriela, Ksawery, Jakub.

Further examples of unusual diminutives or nicknames

- Cesia (Cecylia),
- Nunek (Janusz),
- Neś (Jan),
- Stalek (Stanisław),
- Sławek (Czesław),
- Zońka (Zofia),
- Renia (Renata),
- Hala (Helena),
- Lucia (another Helena),
- Adzio (Waldemar),
- Orcio (Wiktor),
- Zelcia (Roma),
- Andulka (Joanna).

How could I do it? How did I do it?

- There is Morfeusz, a morphological analyzer for Polish by Marcin Woliński, based on *Słownik gramatyczny języka polskiego* (*Grammatical Dictionary of Polish*) — `http://sgjp.pl`.
- It can normalize individual proper nouns to nominative.
- There are also tools for resolution of pronouns.
- But I needed to normalize complex noun phrases, to disambiguate them in context, and to deal with the idiosyncratic diminutives used in my family.
- I developed a set of dedicated Perl scripts from scratch.
- (Defining case inflection was the least problem, in fact.)

The generated index (373 names, 2079 mentions)

Spis postaci

Nazwiska kobiet w tym spisie ułożone są według wszystkich kolejno przybieranych nazwisk, począwszy od nazwiska rodowego, niezależnie od tego, czy owe osoby używały złożonych nazwisk po zamążpójściu. Kursywą odnotowano przydomki, zdrobnienia i warianty ortograficzne imion.

Examples of generated labels

```
W~rodzinnej tradycji przetrwała wiara w~magiczną moc bajki o
\index{Gapcia Kropeczka}%
Gapci Kropeczce, którą Mama musiała wielokrotnie powtarzać na
żądanie majacącej córeczki.

Mój Ojciec, adwokat
\index{\textsc{Bielawski} Jerzy \textit{Jurek}}%
Jerzy Bielawski, zakończył swe trzydziestojednoletnie życie
30~września 1940 roku w~obozie koncentracyjnym Auschwitz.

... i~matka
\index{\textsc{Rodowicz} Jan \textit{Neś Janek Anoda}}%
Nesia\dywiz Janka\dywiz Anody nie żyła już od kilku lat.

Ze łzami w~oczach ,,ciocia''
\index{\textsc{Jarocińska Skibińska Higier} Teresa}%
Teresa musiała jej odmówić, ...
```

More complex examples: Female surnames

```
... on młodszy brat
\index{\textsc{Jaczynowska Bielawska} Paulina \textit{Niusia}}%
Pauliny (Niusi) Bielawskiej, ...

... opowiadające historię rodziny jego babki
\index{\textsc{Jaczynowska Bielawska} Paulina \textit{Niusia}}%
Niusi z~Jaczynowskich Bielawskiej.

\index{\textsc{Jaczynowska Jeziorowska Rauszer} Renata
\textit{Renia}}%
Renata Jaczynowska I-mo voto Jeziorowska, II-voto Rauszerowa
(1916--2006). To ona zdradziła mi parę rodzinnych tajemnic.

Ciotka
\index{\textsc{Bortnowska Rodowicz} Zofia \textit{Zońka}}%
Zońka Rodowiczowa, z~domu Bortnowska, wnuczka
\index{\textsc{Roth Bortnowska} Aniela}%
Anieli ...
```

More complex examples: Coordinated names

```
\index{\textsc{Bielawski} Jerzy \textit{Jurek}}%
\index{\textsc{Bielawska Żakowska} Irena \textit{Irka Iruchna}}%
Jurek i~Irka mówili, że nie mogą zrozumieć, że Wierzchowice już nie
istnieją.

\index{\textsc{Jaczynowski} Jerzy \textit{Jurek}}%
\index{\textsc{Bielawska Żakowska} Irena \textit{Irka Iruchna}}%
Jurek i~Irka byli bardzo kochającym się rodzeństwem. (MISTAKE!)

\index{\textsc{Żakowska Dąbrowska} Hanna \textit{Hanka Hania}}%
\index{\textsc{Żakowska Formańska} Teresa Barbara \textit{Tereska
Terenia}}%
Hania i~Terenia Żakowskie w~czasach peregrynacji (lata 1950-te).
```

pipe.sh

```
#!/bin/sh

cat Bielawscy.tex \
| ../skrypty/generate_index.pl 2>lista_imion.txt \
| ../skrypty/normalize_index.pl | ../skrypty/purge_index.pl \
| ../skrypty/nominative_index.pl 2>lista_problemow.txt \
| ../skrypty/flatten_index.pl \
| ../skrypty/complete_index.pl 2>lista_uzupelnien.txt \
| ../skrypty/beautify_index.pl 2>lista_osob.txt \
| ../skrypty/non_breaking_spaces.pl >Indexed.tex

pdflatex -shell-escape Indexed.tex

pdflatex -shell-escape Indexed.tex

cat lista_uzupelnien.txt | sort | uniq | wc

cat lista_osob.txt | sort | uniq | wc
```

Particular scripts (644 lines in total)

- `generate_index.pl` – transforms all mentions of patterns "G+ (i G+)? (S(-S)*)? (z S)?", where G is a given name and S is a surname, into raw index labels.
- `normalize_index.pl` and `purge_index.pl` – remove redundant spaces from index labels and redundant labels out of the main matter.
- `nominative_index.pl` – turns all labels into nominative, splits coordinated phrases, splits labels into **5 name fields**: *maiden surname, surname, husbands' surnames, given names, and nicknames*.
- `complete_index.pl` – completes missing information in the **5 name fields** by unifying with the nearest matching mentions (forward and backward pass).
- `flatten_index.pl` and `beautify_index.pl` – consolidate the **5 name fields** in the labels into the printable form.

morphology.txt (409 lines)

```
ssa:  *s~ki~kiego~kiemu~kim~cy~kich~kim~kimi~ka~kiej~ką~kie
ssm:   *cz~~a~owi~em~u~owie~ów~om~ami~ach
          ~owa~owej~ową~owe~owych~owym~owymi
          ~ówna~ówny~ównie~ównę~ówną~ówno
          ~ówny~ówien~ównom~ównami~ównach

gmm:Dzid~ek~ka~kowi~kiem~ku     Wiesław
gmm:Francisz~ek~ka~kowi~kiem~ku
gmm:Gen~ek~ka~kowi~kiem~ku      Eugeniusz
gmm:Jac~ek~ka~kowi~kiem~ku
gmm:Jan~ek~ka~kowi~kiem~ku      Jan
gmm:Janusz~ek~ka~kowi~kiem~ku   Janusz
gmm:Jur~ek~ka~kowi~kiem~ku      Jerzy
gmm:Mar~ek~ka~kowi~kiem~ku
gmm:Miet~ek~ka~kowi~kiem~ku     Mieczysław
gmm:Nun~ek~ka~kowi~kiem~ku      Janusz
```

Hacking in Bielawscy.tex (33 names)

```
\mainmatter
\newcommand{\declaration}[1]{}
\declaration{
  Harry (Garri) (Juljewicz) Eduard Lorenzson\
  Erna Paulie Lorenzson Bielawska\
  Irena (Irka) (Iruchna) Bielawska Żakowska\
  Barbara (Basia) (Barbatesa) (Basieńka) Teresa Bielawska Dębowska\
  Jerzyna (Jerzynka) Bielawska Słomczyńska\
  Krystyna (Krysia) (Krzychna) Janina Westerska Bielawska Jackowska\
  Fernande (Nande) Thomas Bielawski\
  Jerzy (Jurek) Bielawski\
  Jan (Neś) (Janek) (Anoda) Rodowicz\
  Helena (Halina) (Hala) Bielawska Barcikowska\
  Samuel Bielawski\
  Teodora Żelazkiewicz Bielawska\
  Hanna (Hania) Westerska Kopczyk\
  ...
}
```

Abstract of this talk by ChatGPT

This paper presents an innovative approach to enriching genealogical publications through the implementation of Named Entity Recognition (NER) for indexing individuals. Inspired by the monumental work of Barbara Bielawska-Dębowska (1937–2020), the presenter embarked on the challenging task of automating the indexing process for her posthumously published book, "Bielawscy – Pięciu w linii prostej."

The presentation commences with an overview of the inherent complexities in indexing familial names, given the nuances of Polish language conventions and the idiosyncrasies of familial nicknames and pseudonyms. Challenges such as normalization, disambiguation, and handling diminutives are elucidated, setting the stage for the proposed computational solution.

The presenter introduces a suite of Perl scripts meticulously crafted for Named Entity Recognition and index generation. Leveraging the Morfeusz morphological analyzer for Polish, the scripts seamlessly normalize complex noun phrases, disambiguate references, and generate index labels semi-automatically. Through a step-by-step demonstration, attendees gain practical insights into the implementation and usage of each script in the indexing pipeline.

My own abstract

We introduce a framework that reduces the burden of making an index of persons for LaTeX files in Polish. Applying the standard indexing mechanism of LaTeX, we augment it with a Perl-based pipeline of scripts that performs named entity recognition (NER). Our work is motivated by the posthumous publication of Barbara Bielawska-Dębowska's (1937–2020) final book, *Bielawscy – Pięciu w linii prostej*, where we applied this pipeline.

The functionality of the NER pipeline is to insert LaTeX index labels into the raw text file. We had to deal with problems such as detection of mentions, normalization of labels, and completion of missing information. Because of high idiosyncrasy of nicknames used by Barbara Bielawska-Dębowska, we did not use Morfeusz, an available morphological analyzer for Polish.

Instead of this, from scratch, we developed a set of dedicated Perl scripts that implement a needed subset of Polish morphology and syntax of personal proper names. Our scripts deal with simple coordination of names and a more complex syntax of female surnames. The normalization of labels involves adding missing information by finding the nearest mentions that can be unified in all five name fields: maiden surname, surname, husbands' surnames, given names, and nicknames.

The pipeline was developed specifically for Barbara Bielawska-Dębowska's book. Not being publically available, it can be shared on request.

Po polsku (tłumaczenie ChatGPT + korekta)

Przedstawiamy narzędzie, które redukuje trudność tworzenia indeksu osób w plikach LaTeX-owych w języku polskim. Wykorzystując standardowy mechanizm indeksowania LaTeX-u, wzbogacamy go o potok skryptów napisanych w Perlu, który wykonuje rozpoznawanie jednostek nazewniczych (named entity recognition, NER). Nasza praca jest motywowana pośmiertnym wydaniem ostatniej książki Barbary Bielawskiej-Dębowskiej (1937–2020), *Bielawscy – Pięciu w linii prostej*, w której zastosowaliśmy ten potok.

Funkcjonalność potoku NER polega na wstawianiu etykiet indeksu LaTeX-owego do pliku z surowym tekstem. Poradziliśmy sobie z problemami takimi jak wykrywanie wzmianek, normalizacja etykiet oraz uzupełnianie brakujących informacji. Ze względu na znaczną idiosynkrazję zdrobnień używanych przez Barbarę Bielawską-Dębowską, nie skorzystaliśmy z Morfeusza, dostępnego analizatora morfologicznego dla języka polskiego.

Zamiast tego, opracowaliśmy od podstaw zestaw dedykowanych skryptów w Perlu, które implementują potrzebny podzbiór morfologii polskiej oraz składni nazw osobowych. Nasze skrypty radzą sobie zarówno z prostą koordynacją imion, jak i z bardziej złożoną składnią nazwisk żeńskich. Normalizacja etykiet polega na dodawaniu brakujących informacji poprzez znalezienie najbliższych wzmianek, które można zunifikować we wszystkich pięciu polach nazw osobowych: nazwisku panieńskim, nazwisku, nazwiskach mężów, imionach oraz pseudonimach.

Potok nasz został opracowany specjalnie na użytek składu książki Barbary Bielawskiej-Dębowskiej. Nie jest on publicznie dostępny, ale może być udostępniony na życzenie.