Ilościowa analiza sieci postaci w polskich powieściach XIX i XX wieku

Marek Kubis

Uniwersytet im. Adama Mickiewicza w Poznaniu
mkubis@amu.edu.pl

12 kwietnia 2021

Wyniki pochodzą z pracy *Quantitative analysis of character networks in Polish 19th- and 20th-century novels*, Digital Scholarship in the Humanities, 2021;
https://doi.org/10.1093/llc/fqab012

Sieci postaci

Sieci postaci

- Postacie są wierzchołkami grafu.
- Dwie postacie są połączone krawędzią, jeżeli uczestniczą w tej samej konwersacji.

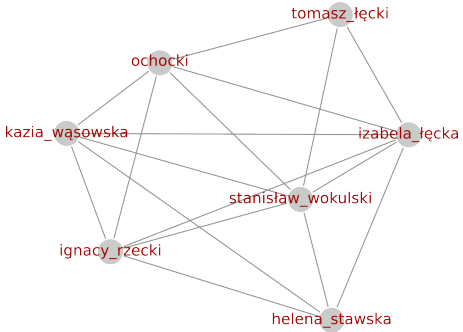


Figure 1: Główne postacie *Lalki* B. Prusa

Sieci postaci

— Pryncypał jeszcze nie wraca , nie miał pan listu ? — **spytał** Klejn .

— Spodziewam się go w połowie marca , najdalej za miesiąc .

— Jeżeli go nie zatrzyma nowa wojna .

— Staś . . . Pan Wokulski — **poprawił się** Rzecki — pisze mi , że wojny nie będzie .

— Kursa jednak spadają , a przed chwilą czytałem , że flota angielska wpłynęła na Dardanele .

— To nic , wojny nie będzie . Zresztą — **westchnął** pan Ignacy — co nas obchodzi wojna , w której nie przyjmie udziału Bonaparte .

— Bonapartowie skończyli już karierę .

— Doprawdy ? . . . — **uśmiechnął się ironicznie** pan Ignacy . — A na czyjąż korzyść MacMahon z Ducrotem układali w styczniu zamach stanu ? . . . Wierz mi , panie Klejn , bonapartyzm to potęga ! . . .

— Jest większa od niej .

— Jaka ? — **oburzył się** pan Ignacy . — Może republika z Gambettą ? . . . Może Bismarck ? . . .

— Socjalizm . . . — szepnął mizerny subiekt kryjąc się za porcelanę .

Pan Ignacy mocniej zasadził binokle i podniósł się na swym fotelu , jakby pragnąc jednym zamachem obalić nową teorię , która przeciwstawiała się jego poglądom , lecz poplątało mu szyki wejście drugiego subiekta z brodą .

Sieci postaci

Prace powiązane

- character networks (Moretti i in.; 2013)

  (postacie wymieniają między sobą akty mowy)
- **conversation networks** (Elson i in.; 2010)

  (postacie uczestniczą w tej samej konwersacji)
- character-localization networks (Lee and Yeung; 2012)

  (postacie pojawiają się w tej samej lokalizacji)

Sieci postaci

Prace powiązane

- interaction networks (Jayannavar i in.; 2015)

  (postacie uczestniczą w tym samym zdarzeniu)

- observation networks (Jayannavar i in.; 2015)

  (jedna z postaci jest świadoma obecności drugiej, ale relacja nie jest symetryczna)

Analiza sieci postaci

Motywacja

- Ilościowa walidacja jakościowych hipotez pojawiających się w teorii literatury (trudne; por. Elson i in. 2010 oraz Jayannavar i in. 2015)
- Analiza zmian w strukturze sieci postaci na przestrzeni czasu (ta praca)
- ...

Analiza sieci postaci

Cel badań

Analiza systematycznych różnic pomiędzy sieciami postaci indukowanymi z prozy polskiej (głównie) drugiej połowy XIX i pierwszej połowy XX wieku w odniesieniu do zgromadzonego korpusu.

Analiza sieci postaci

Cele poboczne

1. Stworzenie automatycznej procedury prowadzącej od korpusu powieści do zbioru testowalnych hipotez.
2. Odtworzenie obserwacji Elsona i in. (2010) i Jayannavara i in. (2015) na bardziej wymagającym korpusie – większym o rząd wielkości i zbudowanym dla innego języka.

Analiza sieci postaci

Procedura

1. Zgromadzenie korpusu powieści
2. Indukcja sieci postaci
3. Formułowanie i testowanie hipotez dot. struktury indukowanych sieci

Korpus powieści

Korpus powieści

Źródła

- źródła – biblioteki cyfrowe:
  - wolnelektury.pl: 230 (obecnie 254) powieści
  - polona.pl: ok. 3000 (obecnie ponad 10000) zdigitalizowanych woluminów zawierających warstwę OCR.
- głównie druga połowa XIX-ego i pierwsza połowa XX-ego wieku
- język: polski

Korpus powieści

Problemy

Korpus pozyskany z bibliotek cyfrowych jest zaszumiony i niezrównoważony.

- zawiera wiele edycji tego samego dzieła
- zawiera zarówno edycje jednotomowe jak i wielotomowe
- większość powieści jest dostępna wyłącznie w formie nieredagowanego tekstu pochodzącego z OCR
- zawiera zarówno teksty polskie jak i tłumaczenia literatury obcojęzycznej

Korpus powieści

Równoważenie korpusu

1. Edycje wielotomowe zostały połączone w całość

   Uzyskano w ten sposób ok. 2500 (obecnie 7000) kompletnych tekstów

2. Do dalszych prac wybrano dokładnie jedną edycję każdej powieści

   - priorytet otrzymały najnowsze edycje nie-OCR
   - do wyznaczenia właściwej edycji wykorzystano metadane z katalogu Biblioteki Narodowej
   - uzyskano 1555 (obecnie 4400) unikalnych tekstów

3. Z korpusu wyłączono tłumaczenia powieści obcojęzycznych i ograniczono przedział czasowy do lat 1800–1945. Ostatecznie w korpusie znalazły się:

   - 392 (obecnie 1010) powieści z XIX-ego wieku
   - 538 (obecnie 1296) powieści z XX-ego wieku

Korpus powieści

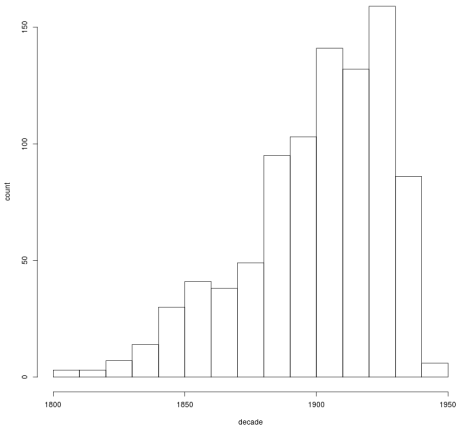
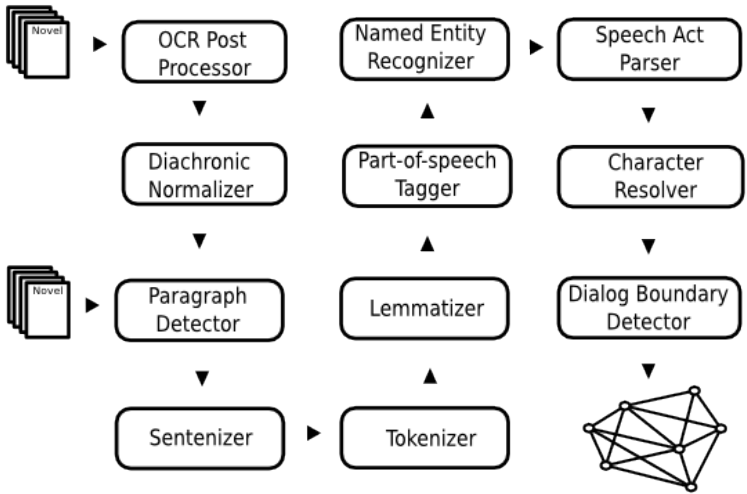


Figure 2: Zawartość korpusu w podziale na dekady

Indukcja sieci

Indukcja sieci

Potok annotacyjny



Indukcja sieci

Potok annotacyjny

par_id	word	lemma	pos	feat	ner	dip	chr	bnd
ch01:004	A	a	conj		0	utt		B
ch01:004	która	który	adj	sg:nom:f	0	utt		B
ch01:004	to	to	pred		0	utt		B
ch01:004	butelka	butelka	subst	sg:nom:f	0	utt		B
ch01:004	?	?	interp		0	utt		B
ch01:005	-	-	interp		0	utt		I
ch01:005	Szósta	szósty	adj	sg:nom:f	0	utt		I
ch01:005	,	,	interp		0	utt		I
ch01:005	panie	pan	subst	sg:voc:m1	0	utt		I
ch01:005	radco	radca	subst	sg:voc:m1	0	utt		I
ch01:005	.	.	interp		0	utt		I
ch01:005	Służę	służyć	fin	sg:pri:imperf	0	utt		I
ch01:005	piorunem	piorun	subst	sg:inst:m3	0	utt		I
ch01:005	!	!	interp		0	utt		I
ch01:005	...	...	subst		0	utt		I
ch01:005	-	-	interp		0			I
ch01:005	odpowiadał	odpowiadać	praet	sg:m1:imperf	0	action		I
ch01:005	Józio	józio	subst	sg:nom:m1	B-persName	speaker.name	józio	I
ch01:005	.	.	interp		0			I

Parser aktów mowy – powierzchniowy parser służący do identyfikacji mówców.

*— Jaka? — **oburzył się** pan Ignacy. — Może republika z Gambettą?... Może Bismarck?...*

```
- Jaka/utt ?/utt
- oburzył/action się/action pan/speaker Ignacy/speaker .
- Może republika z Gambettą ? ... Może Bismarck ? ...
```

Indukcja sieci

Parser aktów mowy

— Pryncypał jeszcze nie wraca , nie miał pan listu ? — **spytał** Klejn .

— Spodziewam się go w połowie marca , najdalej za miesiąc .

— Jeżeli go nie zatrzyma nowa wojna .

— Staś . . . Pan Wokulski — **poprawił się** Rzecki — pisze mi , że wojny nie będzie .

— Kursa jednak spadają , a przed chwilą czytałem , że flota angielska wpłynęła na Dardanele .

— To nic , wojny nie będzie . Zresztą — **westchnął** pan Ignacy — co nas obchodzi wojna , w której nie przyjmie udziału Bonaparte .

— Bonapartowie skończyli już karierę .

— Doprawdy ? . . . — **uśmiechnął się ironicznie** pan Ignacy . — A na czyjąż korzyść MacMahon z Ducrotem układali w styczniu zamach stanu ? . . . Wierz mi , panie Klejn , bonapartyzm to potęga ! . . .

— Jest większa od niej .

— Jaka ? — **oburzył się** pan Ignacy . — Może republika z Gambettą ? . . . Może Bismarck ? . . .

— Socjalizm . . . — szepnął mizerny subiekt kryjąc się za porcelanę .

Pan Ignacy mocniej zasadził binokle i podniósł się na swym fotelu , jakby pragnąc jednym zamachem obalić nową teorię , która przeciwstawiała się jego poglądom , lecz poplątało mu szyki wejście drugiego subiekta z brodą .

Indukcja sieci

Ujednoznacznianie nazw postaci

- Problem: Postać literacka może występować w tekście pod różnymi nazwami.

  — **Staś** ... Pan **Wokulski** — poprawił się **Rzecki** — pisze mi , że wojny nie będzie .

  — Doprawdy ? ... — uśmiechnął się ironicznie pan **Ignacy** .

- Cel: Przypisać wszystkim nazwom odnoszącym się do tej samej postaci wspólny identyfikator

  — **Staś/1** ... Pan **Wokulski/1** — poprawił się **Rzecki/2** — pisze mi , że wojny nie będzie .

  — Doprawdy ? ... — uśmiechnął się ironicznie pan **Ignacy/2** .

Indukcja sieci

Ujednoznacznianie nazw postaci: związki z innymi problemami

1. Named Entity Recognition

— **Staś/B** ... Pan **Wokulski/B** — poprawił się **Rzecki/B** — pisze mi , że wojny nie będzie .

2. Character name linking/resolution/identification

— **Staś/1** ... Pan **Wokulski/1** — poprawił się **Rzecki/2** — pisze mi , że wojny nie będzie .

3. Coreference resolution

— **Staś/1** ... Pan **Wokulski/1** — poprawił się **Rzecki/2** — pisze **mi/2** , że wojny nie będzie .

Indukcja sieci

Ujednoznacznianie nazw postaci: nienadzorowany algorytm grafowy

1. Utwórz wierzchołek w grafie dla każdej nazwy postaci występującej w tekście.
2. Połącz odniesienia, jeżeli mają te same lematy.

   Stanisława Wokulskiego – Stanisławem Wokulskim
3. Połącz zdrobnienia z formami podstawowymi.

   Stanisław – Stasiek
4. Połącz skróty z rozwinięciami podstawowymi.

   S. Wokulski – Stanisław Wokulski
5. Usuń obserwacje odstające (singletony, niemowy itp.)

Indukcja sieci

Ujednoznacznianie nazw postaci: nienadzorowany algorytm grafowy – cd.

6. Połącz afiksy z najbliższymi im odniesieniami, które je obejmują.

   Stanisław – Stanisław Wokulski

7. Rozłącz niejednoznaczne odniesienia.

   (skorzystaj z grafu ilorazowego względem lematów)

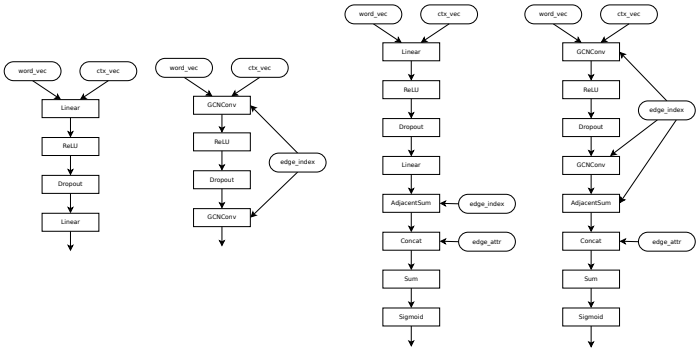
   Franc – Franc Mincel -/- Mincel -/- Jan Mincel – Jan

8. Wyznacz składowe spójne grafu.

   Stanisław – Stasiek – Wokulski – Stanisław Wokulski

Indukcja sieci

Ujednoznacznianie nazw postaci: nadzorowane algorytmy grafowe



Geometric Deep Learning Models for Linking Character Names in Novels (Kubis, 2020)
https://www.aclweb.org/anthology/2020.latechclfl-1.15.pdf

Indukcja sieci

Wykrywanie granic konwersacji

- Granice dialogu są identyfikowane na podstawie odległości (w tekście) pomiędzy sparsowanymi aktami mowy.
- Dla każdych dwóch postaci uczestniczących w konwersacji jest tworzone połączenie w wynikowej sieci.

Wyniki

Własności sieci v. liczba postaci

Elson i in. (2010) oraz Jayannavar i in. (2015) ustalili, że liczba postaci koreluje z poszczególnymi własnościami indukowanych sieci.

Tę samą prawidłowość (co nie powinno zaskakiwać) można zaobserwować w przypadku polskich powieści.

Table 1: Korelacja własności sieci z liczbą postaci

Własność	XIX wiek	XX wiek	Wszystkie
liczba konwersacji	0.74	0.73	0.73
liczba krawędzi	0.84	0.88	0.86
średni stopień wierzchołka	0.35	0.54	0.44
liczba społeczności	0.83	0.83	0.83
wielkość społeczności	0.40	0.46	0.43

Własności sieci v. stulecie

Table 2: Wartości średnie własności sieci w podziale na wiek

Własność	XIX wiek	XX wiek	Wszystkie
liczba postaci	36.87	31.48	33.75
liczba konwersacji	182.60	196.27	190.50
liczba krawędzi	89.56	73.16	80.09
średni stopień wierzchołka	2.35	2.23	2.28
liczba społeczności	5.13	4.35	4.68
wielkość społeczności	5.83	5.56	5.68

Własności sieci v. stulecie

Obserwacje

- Liczba postaci, konwersacji, krawędzi oraz liczba społeczności wydaje się różnić pomiędzy stuleciami.
- Własności sieci **nie pochodzą** z rozkładu normalnego (test Shapiro-Wilka).
- Skorzystałem z nieparametrycznego testu Manna-Whitneya żeby sprawdzić hipotezę:

  It is equally likely that a randomly selected novel from the 19th century subcorpus has a lower or higher value of the network metric being tested than a randomly selected novel from the 20th century subcorpus.

Table 3: Wyniki testu Manna-Whitneya

Własność	mediana XIX	mediana XX	p-value	0.95 przed. ufności
liczba postaci	30.0	26.0	**0.00098**	[2.00, 6.00]
liczba konwersacji	132.0	161.5	**0.00671**	[-34.0, -6.0]
liczba krawędzi	54.0	43.0	**0.00370**	[3.00, 15.00]
średni stopień wierzchołka	1.9	1.9	0.68892	[-0.14, 0.21]
liczba społeczności	4.0	4.0	**0.00020**	[0.000068, 1.00]
wielkość społeczności	5.2	5.0	0.26845	[-0.117, 0.49]

Własności sieci v. stulecie

Wnioski

- Hipotezy zostały odrzucone w przypadku liczby postaci, liczby konwersacji, liczby krawędzi oraz liczby społeczności,
- Wskazuje to na to, że (przynajmniej w odniesieniu do zgromadzonego korpusu i wybranej metody indukcji sieci) proza pierwszej połowy XX wieku jest bogatsza w dialogi oraz koncentruje się na mniejszej liczbie postaci.

Pytania?

Bibliografia

Elson, D. K., Dames, N. and McKeown, K. R. (2010). Extracting social networks from literary fiction. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 138–147.

Jayannavar, P., Agarwal, A., Ju, M. and Rambow, O. (2015). Validating Literary Theories Using Automatic Social Network Extraction. Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature, pp. 32–41

Kubis, M. (2020). Geometric Deep Learning Models for Linking Character Names in Novels. Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. pp. 127–132

Bibliografia (cd.)

Lee, J. and Yeung, Y. (2012). Extracting Networks of People and Places from Literary Texts. Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation, pp. 209–218

Moretti, F. (2013). Network Theory, Plot Analysis. Distant Reading. London: Verso.