Segmentation of Logical Units in Text

Ing. Martin Kostelník

Supervisor: Ing. Karel Beneš



What is TextBite?

A deep learning tool for extraction of articles, news, dictionary entries, book segments, etc. from historical documents.

TextBite enhances search capabilites in digitized documents used by librarians and scientists.

01 / Data

A custom dataset was created and labeled based upon data from Czech Digital Library.

In total, 4044 pages were manually selected and labeled with the help of librarians and other students.

02 / Methodology

- 1. Regions of interest are detected using the YOLOv8 detector.
- A complete graph is constructed from the detected regions. Each node is represented by both geometric and language features.
- 3. The graph is fed into a graph neural network, which works as a binary edge classifier. Regions connected by a positive edge are merged.

04 / Evaluation and Results

The problem is defined as a line clustering problem. Each cluster represents a coherent text segment.

The performance is evaluated using the V-Measure clustering metric (higher is better).

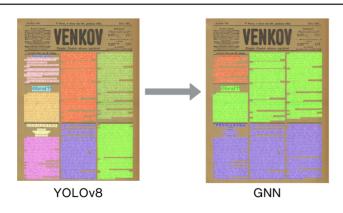
The baseline solution is based on imposing geometric constraints on the consecusion of text lines, with additional distance based splitting or language model splitting.

١/	N/		cı	ıre
v-	IV	2	SI	$\Pi \hookrightarrow$

Method	Books	Dictionaries	Periodical	Time [s]
Baseline	19.72	40.11	52.20	0.16
Baseline+Dist	49.77	64.67	77.40	0.16
Baseline+LM	44.32	77.19	70.66	1.50
YOLOv8	65.33	93.36	83.93	0.14
YOLOv8+GNN	77.93	95.79	90.23	0.66
			On NVIDIA R	TX A6000

05 / Showcase

03 / Graph Neural Network Effect







	zlatosrstný-zlaťou
valatovstný az mořící látrou ort. Zlatovstný při Dilei jen pra nie vice. Theor. valatost, -i r. odminost něleho zlatého. Nepři- penini so jeditivovní prelita ve zlatím orná- til starovštvoné zlatí tek, jemží re Zlicí kla- viti starovštvoné zlatí tek, jemží re Zlicí kla- ziné Prahy? Deny jasu temě, kelitel, očes- zně. kláze.	a počátku II. saletli, Manuali, U oboru a tapackába a kawetepeckába je ládeset te takán jesnycké jako u poslacovačí, Nio p. totepecká blížna Másu ze slepišto atřene k ríbo dolytku Jenu, i katotepecké, il a řemenčně čímost zab jicí se opjávim slate, stříbu, okumicí o k věck slám se vojní traké liste.
"ziatostoloc, -lee m. trán. Jakýsi divoký po-	zlatotisk, -a m. tisk prosedený v barod n
oud veli mi, abych tato modiu a muroovaného	dobicí zlato. Ornament je proveden bo
slatostelce povalil, Zey, zkatostrumý alj. Ma. mojící zleté strany. Já jem v poulit harla slatostrumů, která chvříc se, kdyl vzejde luza. Veh. D pocházylící ze zlo- tých stran. Stylelat dole, kterák zaměl zpěv stříbročistý, zlatostrumý. Tab.	v zlatotiku. Tobol. falotikaný aj. šia. ntkoný ze zlate, zlatem čhásoný. Lema klosla v mělkěř, zlateklaně šišře. Zev. Vchod [lešnice] závoj šlenhý, zlaný kryje, vsa. Jako zlatotkaný kryje nako zlate závoje před pokrývá pocině její (Kunžitick) boje. Vec. Al za blížo klenně těje, apoší v jeve. Al za hlížo klenně těje, apoší v jeve. Al za hlížo klenně těje, apoší v jeve.
"zintostřebec, -bce m. hlodoč zlate. Hltavé zlatostřebce a houby nenasytné jsem napájel dary své štědrosti. Bost.	padne s výše alatotkaný štěstí sen. Heyd. lumá, kouzelná, sní v září zlatotkané. Ltd. so
zlatostříbený saj, zlatý a stříbvej. Kutorovy výlevy i s vavřísovým věncem a zlatostříbený- mi stulkami octly se ve vaku umělkyně. Patel	*zlatotoký sij. (o tehaticě) mající barou alatova, Mlovém o padesátí lahvich výteže obnivého, zlatotokého, původního mazad páleného to dalmatikého výta, tod.
*ziatolatý ad. mající zloté laty, jsoucí ve zla- tých lotech. Tam, hle, v dáli Praha stověžatá, jeho (Karla IV.) dcera září zlatolatá, O.Core.	zlatotrůmný adj. bás. mající zlatý tráv. S [Diem] zlatotrůmná šla Héré. vas.
ziatošodivě adv. zletě a ledivě. Vrcholck věle alabě se povyjaznil a její okraj zbarvil se zlato- ledivě. Kom.	*ziatotřebecký ad. alchymistický. Na to zámku umční ziatotřebeckého vyhledávali.
sidne. Rom.	*zlatotřebství, -i n. alchymie. Lepeni l
zlatošodý azj. Jedý s múdechem do zlatovu.	z rina jakýmu bílým práklem posypino
Druž oliř libula se a druhem tak vroucoš, pyl	těch dob, co se pan Odolan zabýval zlatot
až zlatošedý nes' a jehošdek se na jehošdy. Heyš.	stvím. Těb.
ziatolpičký až, aud ilutopásec ziatolpičký drah	*zlatotřesotný "dj. zlatě se třpytící. Jak o
lvouhi z teledi Cerambycidar, Plagionstus de-	říly [na vánodním stroměu] zlatotřesotné h
tritus.	dy a zlaté vlasy! Woja.
zlatoštitký sij. 2001. licoltitkië zlatoltitký drah Ivouhů z čeledí Buprestidae, Eurythyrea scutel- toris.	*zlatotřpyt, -u m. třpyt zleta, zlatý lesk. pešadné makovice, vycídění k falešnému ú řpytu, nutno srasití. Mach. (Skřívánčí pl nad rožem se chvěte v uzrářých klasů zl
zlatoštitý sit, zeel, polník zlatoštitý brouk z če-	tlpyt. F.S.Prod.
ledi Buprestidoe, Agrilus viridis.	zlatotipytný atj. šta. třpytící se jako z
ziatotečký ads. sod. vyklezulec zlatotečký od-	Cetné pokoje sotva stačily hostům vznele:
rálda brouků z čeledí Byrrhidae, Seminolas pi-	již nosivali slatotlpytnou uniforma pr. Z
tula var. awatopunctetsu.	co děvče jedlo, srovnala mu paní vlasv sli
zlatotočný sás. sod. zlatojamec zlatotelný drah	hřebenem, a tu zkadeřila se ta zlatotřpy
brouků z čeledi Bujrestidae, Chrysobothris chry-	milá trářička. Vesk. Zost hubokaz zlatotřp
costizme.	drah brouhli z čeledi Giridse, Gis setiger.
zlatotělka, vy t. 2011. rod much z čeledí Denii-	*ziatotvorce, zlatotvůrce, -c m (dosu
dae, Leskia; rod much z čeledí Syrphidae, Chry-	tuárce, výrobce zlate, alchymista. Pověst p
sochlanez.	Jila také na nelivé tělo pozdního alatetv
ristotopoc, -pce m. řemeninik, precovnik robý-	remantickou kytičku. Lit. sov. Jako véveda
vojicí se tepinim siete. Tv. (odpadky zlata)	chovali Karneela Zidé v poespou. – bylť sy
divá tatik maž za tu práci. Střádka ni je, a si	jejich kmese, testo zlatetvorce. Soch.
jich bude komička natlačená, prodám je buď	zlatouci stj. sop. (o boroš) (přijemně, he
jich sude konvicta natucena, prodam je bud	zistroven 25, 256. (o serve) pryserve, ne
idanskovi nebo oškierešmu slatotepci. Kas. Nes.	zistrovený do zlatovar. V člíších zašunžilo a
Zlatotepec podzim pemalu pokrývá usinající	perillo se zlatovať vino, třeme. D = 2004. z
zemi kyprou podatkou opadávajících listů. Zper-	zlatovať pravé, čisté zlato, Elizounico, na
volsí.	pod stolní doskou, zlato zlatovací, nevidané.
zlatotepecký aj, k zletotepec, Remo [bylo] sidlem dilny, která vytvočila nejvýmanodjil kus zlatotepecké práce v antipendiu basilejském	zlatoučký ult. esp. (příjemně, kezky) zbar do zlatova, zlatové řesklý. Staničko v záli