

Upraszczanie tekstu przez podmiany wielowyrzowe

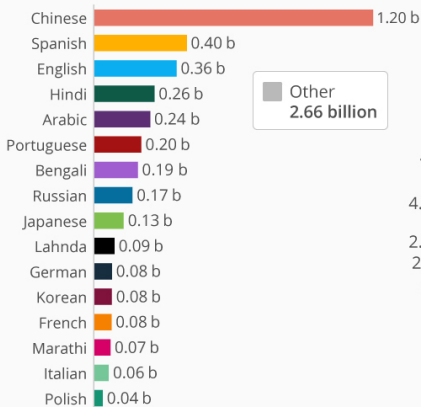
Piotr Przybyła
`piotr.przybyla@ipipan.waw.pl`

Instytut Podstaw Informatyki PAN

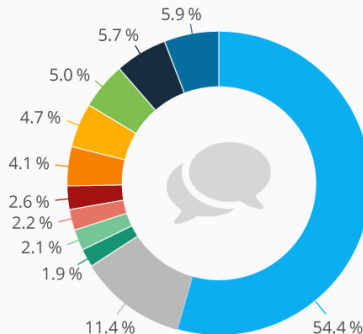
17 grudnia 2020

Languages Most Used on the Web vs. IRL

Number of first-language speakers (estimates in billions)

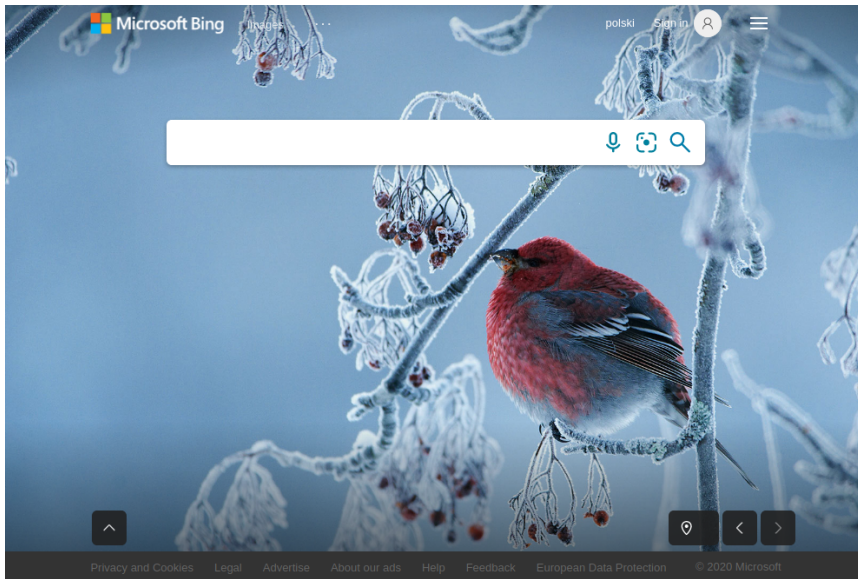


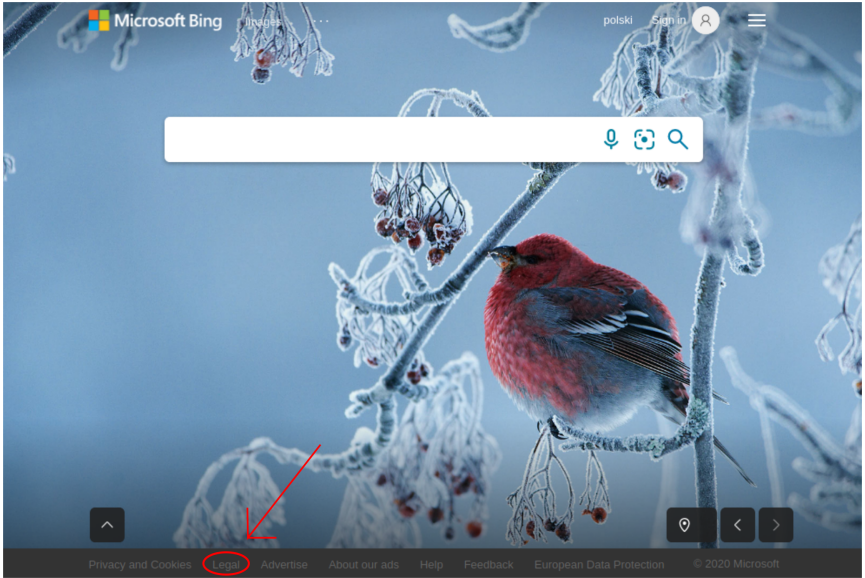
Percentage of websites using various content languages*



* top 10 million websites

Sources: w3techs.com, Ethnologue





If you have any basis for recovering damages (including breach of these Terms), you agree that your exclusive remedy is to recover, from Microsoft or any affiliates, resellers, distributors, Third-Party Apps and Services providers, and vendors, direct damages up to an amount equal to your Services fee for the month during which the loss or breach occurred (or up to \$10.00 if the Services are free). You can't recover any other damages or losses, including direct, consequential, lost profits, special, indirect, incidental, or punitive. These limitations and exclusions apply even if this remedy doesn't fully compensate you for any losses or fails of its essential purpose or if we knew or should have known about the possibility of the damages. To the maximum extent permitted by law, these limitations and exclusions apply to anything or any claims related to these Terms, the Services, or the software related to the Services.

*If you have any basis for recovering damages (including breach of these Terms), you agree that your exclusive remedy is to recover, from Microsoft or any affiliates, resellers, distributors, Third-Party Apps and Services providers, and vendors, direct damages up to an amount equal to your Services fee for the month during which the loss or breach occurred (or up to **\$10.00** if the Services are free). You can't recover any other damages or losses, including direct, consequential, lost profits, special, indirect, incidental, or punitive. These limitations and exclusions apply even if this remedy doesn't fully compensate you for any losses or fails of its essential purpose or if we knew or should have known about the possibility of the damages. To the maximum extent permitted by law, these limitations and exclusions apply to anything or any claims related to these Terms, the Services, or the software related to the Services.*

Upraszczenie tekstu

Manchester
Metropolitan
University

Modyfikacja danego tekstu tak, by uczynić go prostszym do zrozumienia przy zachowaniu znaczenia.

Zastosowania:

- nie-rodzimi użytkownicy języka [Paetzold and Specia, 2016],
- nie-eksperti w dziedzinie tekstu [Cao et al., 2020],
- osoby z zaburzeniami funkcji językowych np. afazją [Shardlow, 2015],

Przykład:

The current lecture appeals to me. → I like this talk.

powiązane: *style transfer*

Dwa podejścia



Manchester
Metropolitan
University



Machine translation (MT): podejście jak do tłumaczenia [Zhu et al., 2010]:

- wymagane duże zbiory danych,
- nieograniczone modyfikacje.

Lexical simplification (LS): podmiana pojedynczych słów [Shardlow, 2014]:

1. Wykrywanie skomplikowanych słów (*Complex Word Identification*, CWI), np. *lecture*,
2. Generowanie synonimów, np. *talk*, *presentation*, *speech*,
3. Wybór najłatwiejszego słowa, np. *talk*.

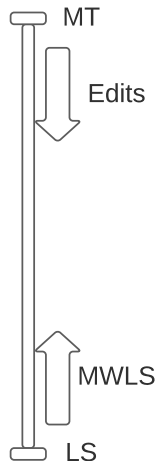
Poziomy upraszczania



Manchester
Metropolitan
University



- żadna ze skrajności nie odzwierciedla sposobu tworzenia uproszczeń przez ludzi,
- próby osiągnięcia złotego środka z dwóch stron:
 - tworzenie zasobów [Alva-Manchego et al., 2020] i metod [Dong et al., 2019] reprezentacji rozwiązań MT przez serię modyfikacji,
 - rozszerzanie schematu LS o bogatsze zmiany (MWLS).



Wkład



Manchester
Metropolitan
University



- Zdefiniowanie zadania upraszczania tekstu przez podmiiany wielowyrazowe (MWLS),
- Przygotowanie w drodze *crowdsourcingu* zbioru podmian wykonanych przez ludzi (MWLS1),
- Zaprojektowanie i implementacja nienadzorowanego rozwiązania automatycznie generującego uproszczenia (*Plainifier*).

Praca zespołowa z MMU (UK)

P. Przybyła, M. Shardlow, "Multi-Word Lexical Simplification," w *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, Barcelona, Hiszpania, 2020.

<https://www.aclweb.org/anthology/2020.coling-main.123/>



Zadanie

Zadanie

Manchester
Metropolitan
University

Upraszczanie tekstu przez podmiiany wielowyrazowe (*Multi-Word Lexical Simplification*, MWLS) polega na zastąpieniu wskazanego fragmentu zdania (1-N słów) innym fragmentem (1-N słów), tak, aby zmodyfikowane zdanie:

1. zachowało to samo znaczenie, co oryginał,
2. było poprawne w języku naturalnym,
3. było łatwiejsze do zrozumienia niż oryginał.

*Our goal is to **provide support** to the students.*

→ *give help to, support, help*

Uwagi



Manchester
Metropolitan
University



- w naszych eksperymentach $N=3$, język naturalny=Angielski,
- *fragment* to dowolna sekwencja słów, niekoniecznie wyrażenie wielowyrazowe (MWE),
- wybór fragmentu nie jest częścią zadania MWLS, lecz CWI.

Przykłady

Manchester
Metropolitan
University

Sentence	Replacement
The main difference in the two lines essentially resides in the strength of the promoter.	is basically
Thus says Yahweh of Armies, "They shall thoroughly glean the remnant of Israel. Turn again your hand as a grape gatherer into the baskets."	will gather
I support Ms Lulling's recommendations that the national systems should recognise the importance of protecting self-employed workers, and we should stand against all forms of discrimination , but (...).	bias and unfairness
Other potentially biologically relevant substrates include cholecystokinin and possibly other neuro-peptides [21].	relevant
A man's foes will be those of his own household.	enemies



Zbiór danych MWLS1

Źródła tekstu



Manchester
Metropolitan
University



Trzy źródła tekstu z różnymi rodzajami trudności:

- BIBLE: *World English Bible* (współczesne tłumaczenie) z korpusu równoległego [Christodouloupoulos and Steedman, 2015],
- EUROPARL: Protokoły parlamentu europejskiego z korpusu równoległego [Koehn, 2005],
- BIOMED: Treść artykułów naukowych z dziedzin biomedycznych w korpusie CRAFT [Bada et al., 2012].

Generowanie zadań



Manchester
Metropolitan
University



1. Pobranie losowo wybranych 10,000 zdań (wersetów) z każdego źródła,
2. Określenie trudności każdego słowa używając neuronowego rozwiązania CWI [Gooding and Kochmar, 2019],
3. Dla $N=3, 2, 1$:
 - 3.1 Dla źródeł BIBLE, EUROPARL, BIOMED:
 - 3.1.1 Wybierz zdania o największej średniej trudności fragmentów o długości N ,
 - 3.1.2 Usuń te zdania ze zbioru

Przy określaniu trudności ignorowane nazwy własne, np. *Aaron*, i bardzo rzadkie słowa (w mniej niż 2% dokumentów z Google Books), np. *diacylglycerol*.

Przykłady



Manchester
Metropolitan
University



- BIBLE

- *Therefore Job opens his mouth with empty talk, and he **multiplies** words without knowledge.*
- *When Aaron lights the lamps at evening, he shall burn it, a perpetual incense before Yahweh **throughout your generations** .*

- EUROPARL

- *Mr President, the German Government's initiative and the report we are debating are not **innocently technical**, as they are being made out to be.*
- *But, as has been said, we must also listen to the other 26 countries, their results and also concerns that may be **expressed** during their ratification.*

- BIOMED

- ***Basal urine concentration** in mutant mice is about 161 mOsm, compared to about 1,293 mOsm in wild-type mice ($p < 0.001$).*
- *We also examined the **potential impact** of the transgene on the expression of the neighboring genes.*

Crowdsourcing



Manchester
Metropolitan
University



Crowdsourcing z użyciem *Amazon Mechanical Turk*:

- każde zadanie przekazane 5 wykonawcom,
- wykonawcy pochodzą z UK, USA, Kanady, NZ lub Australii i ukończyli min. 1000 zadań z 98% zaakceptowanych wyników,
- oczekiwane 1-3 słów lub '?', gdy zdanie zbyt trudne lub nie da się znaleźć uproszczenia,
- weryfikacja przez ręczne przejrzenie:
 - odrzucone 4.49% wyników,
 - najczęściej (42%) więcej niż 3 słowa,
 - więcej niż 3% odrzuceń dyskwalifikuje wykonawcę.

Instrukcja



Manchester
Metropolitan
University



Make a given sentence simpler (easier to read and understand) by replacing the highlighted fragment using **1, 2 or 3 words**. Enter your replacement fragment (NOT a whole sentence) in the input field of the form. The modified sentence should:

1. be more easily understandable to less proficient (for example, non-native) speakers of English,
2. keep the general meaning of the original,
3. remain a correct sentence in English and 'read well'.

If you can't understand the original sentence or find a good enough replacement, use a question mark '?' instead.

Please **do not**:

- provide more than one replacement for a fragment,
- use more than 3 or less than 1 words,
- overuse the question mark when a reasonable replacement is possible,
- simply copy the highlighted fragment.

You **can**:

- freely choose a replacement when more than one is possible, for example: *attire* → *dress*, *attire* → *outfit*, *attire* → *clothes*,
- change the grammatical structure of the sentence, for example: *is provided with* → *gets*,
- use a replacement even when its meaning is not exactly the same, but just close enough, for example: *stimulate* → *help*, *numerous sins* → *many crimes*,
- use a replacement with a different number of words than original, for example *drink excessively* → *overdrink*, *drink excessively* → *drink too much*.

The results will be evaluated manually.

Interfejs



Manchester
Metropolitan
University



Replace the **highlighted words** with your own text fragment (up to 3 words long), making the sentence easier to read and understand, while preserving the original meaning:

Original:

*The granting of assistance, residence status and support to find a better alternative for labour and life will secure the cooperation of trafficked persons in **criminal proceedings** and make it easier to punish the perpetrators.*

Simplified:

The granting of assistance, residence status and support to find a better alternative for labour and life will secure the cooperation of trafficked persons in ...

your replacement

... and make it easier to punish the perpetrators.

Zbiór: statystyki



Manchester
Metropolitan
University



W sumie 7059 podmian dla 1462 zdań.

Subset	Sentences	Replacements	Complete	Words	Agreement
BIBLE 1	168	4.9345	95.23%	1.2393	2.1548
BIBLE 2	166	4.8675	89.76%	2.0169	1.3373
BIBLE 3	160	4.9125	92.50%	2.7025	1.2625
EUROPARL 1	166	4.9699	98.19%	1.2614	2.1386
EUROPARL 2	158	4.9241	94.30%	2.0418	1.2025
EUROPARL 3	161	4.8447	87.58%	2.6671	1.0994
BIOMED 1	166	4.8494	87.35%	1.2024	2.0542
BIOMED 2	162	4.7531	82.10%	1.9741	1.2654
BIOMED 3	155	4.3742	57.42%	2.3652	1.0581
BIBLE All	494	4.9049	92.5%	1.9745	1.5911
EUROPARL All	485	4.9134	93.4%	1.9823	1.4887
BIOMED All	483	4.6646	75.98%	1.8344	1.47
All 1	500	4.918	93.60%	1.2344	2.116
All 2	486	4.8477	88.68%	2.0107	1.2695
All 3	476	4.7143	79.41%	2.5807	1.1408
All	1462	4.8283	87.34%	1.9308	1.5171

<https://github.com/piotrrmp/mwls1>



Plainifier

Plainifier

Manchester
Metropolitan
University

Plainifier automatycznie generuje uproszczenia według zasad MWLS. Bazuje na wcześniejszych nienadzorowanych rozwiązaniach [Qiang et al., 2020], rozszerzonych o wielowyrazowość.

Zasada działania:

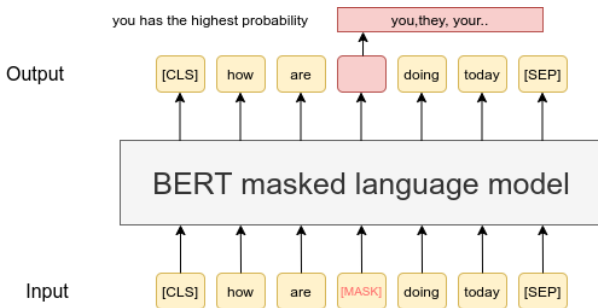
1. Generowanie kandydatów do wypełnienia luki z użyciem modelu językowego (np. BERT),
2. Ocena kandydatów według 3 kryteriów:
 - prawdopodobieństwa (płynności),
 - podobieństwa (zachowania znaczenia),
 - powszechności (łatwości).

<https://github.com/piotrrmp/plainifier>

BERT jako model językowy



Manchester
Metropolitan
University



Source <https://arxiv.org/abs/2003.11562>

TerseBERT: model

Manchester
Metropolitan
University

Zwykły model językowy zwraca prawdopodobieństwo tokenu t^* na podstawie lewego $\mathbf{c}_l = \langle \dots, c_{-2}, c_{-1} \rangle$ i prawego $\mathbf{c}_r = \langle c_1, c_2, \dots \rangle$ kontekstu:

$$LM(t^*|\mathbf{c}_l, \mathbf{c}_r) = \frac{P(\langle \mathbf{c}_l, t^*, \mathbf{c}_r \rangle)}{\sum_t P(\langle \mathbf{c}_l, t, \mathbf{c}_r \rangle)}$$

TerseBERT pozwala określić też prawdopodobieństwo, że $\mathbf{c}_l$ i $\mathbf{c}_r$ następują bezpośrednio po sobie:

$$LM'(\emptyset|\mathbf{c}_l, \mathbf{c}_r) = \frac{P(\langle \mathbf{c}_l, \mathbf{c}_r \rangle)}{P(\langle \mathbf{c}_l, \mathbf{c}_r \rangle) + \sum_t P(\langle \mathbf{c}_l, t, \mathbf{c}_r \rangle)}$$

$$LM'(t^*|\mathbf{c}_l, \mathbf{c}_r) = \frac{P(\langle \mathbf{c}_l, t^*, \mathbf{c}_r \rangle)}{P(\langle \mathbf{c}_l, \mathbf{c}_r \rangle) + \sum_t P(\langle \mathbf{c}_l, t, \mathbf{c}_r \rangle)}$$

TerseBERT

Manchester
Metropolitan
University

1. Zaczynamy od pretrenowanego modelu BERT (BERT-Large, Uncased, WWM),
2. Wznawiamy trenowanie, przy czym 1/3 tokenów [MASK] wstawiona **pomiędzy**, a nie **zamiast** oryginalnych,
3. W tych przypadkach oczekujemy specjalnego tokenu [NONE],
4. Trenujemy 5000 kroków, 128 zdań z Wikipedii w każdym.

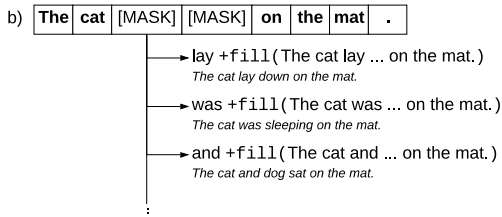
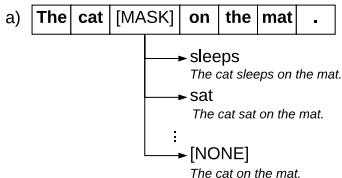
	The	fat	cat	reposed	on	the	mat
P([NONE])	-	0.900	0.208	0.599	-	-	0.005
Token 1	-	black	one	is	-	-	bed
Token 2	-	big	man	lay	-	-	floor

<https://github.com/piotrmp/tersebert>

Generowanie kandydatów

Manchester
Metropolitan
University

fill(The cat ... on the mat.)



Ograniczenia:

- Stop, jeśli w (a) otrzymamy $P([NONE]) > 0.5$
 - Stop, jeśli wygenerowano 3 tokeny,
 - W (b) tylko 16 kandydatów, potem 8, 4, ...,
- + dwustronność, gwarancja oryginalnej ścieżki

Rangowanie kandydatów



Manchester
Metropolitan
University



- Prawdopodobieństwo (P)
 - mierzone przez prawdopodobieństwa z modelu językowego,
 - agregowane przez iloczyn,
- Podobieństwo (S)
 - mierzone przez cosinus wektorów *fastText*,
 - agregowane przez najlepsze wyrównanie,
- Powszechność (F)
 - mierzone przez liczbę dokumentów z wystąpieniami słowa wg *Google Books Ngrams* (skalowane do 0-1),
 - agregowane przez minimum,

$$\text{score}(c) = [P(c)]^{\alpha_1} \times [S(c)]^{\alpha_2} \times [F(c)]^{\alpha_3}$$



Evaluacja

Miary ewaluacyjne

Manchester
Metropolitan
University

Traktując jako zadanie IR, gdzie porównujemy zaproponowane uproszczenia do podanych przez wykonawców (złotych):

- **Prec@5**,
- **NDCG**.

Miary rozluźnione, używane w upraszczaniu tekstu:

- **Potential**: w ilu przypadkach przynajmniej jedno ze złotych uproszczeń obecne na liście rankingowej (gdziekolwiek),
- **BOW@5**: jaka część słów ze złotych uproszczeń zawiera się w zbiorze słów z pierwszych 5 wyników.

Poziomy odniesienia



Manchester
Metropolitan
University



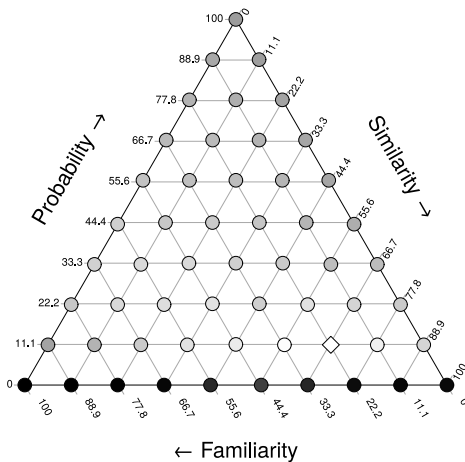
- **Human:** podmiana jednego z wykonawców traktowana jako lista propozycji, pozostałe jako prawidłowe,
- Na bazie zasobów:
 - **SimplePPDB:** poszukuje w bazie *SimplePPDB* [Pavlick and Callison-Burch, 2016] parafraz całego fragmentu, potem według słów. Sortowane według podanego wskaźnika trudności.
 - **WordNet:** analogicznie z użyciem *Princeton WordNet*. Sortowanie według częstości unigramów w *Google Web1T*.

Dostrajanie

Manchester
Metropolitan
University

Dla ustalenia optymalnych wartości $\alpha_1, \alpha_2, \alpha_3$:

1. Wydzielenie 100 zdań ze zbioru ewaluacyjnego,
2. Sprawdzenie 55 kombinacji
 $\alpha_1 = \frac{i}{9}, \alpha_2 = \frac{j}{9}, \alpha_3 = \frac{k}{9}$,
tak, że $i, j, k \in \{0 \dots 9\}, i + j + k = 9$
3. Wybór najlepszej konfiguracji wg NDCG.



Wyniki (NDCG, P@5)



Manchester
Metropolitan
University



Metric	Subset	Plainifier		Human	Baselines	WordNet
		Default	Tuned		SimplePPDB	
NDCG	BIBLE all	0.1505	0.1528	0.1116	0.0307	0.0293
	EUROPARL all	0.1453	0.1575	0.0999	0.0453	0.0267
	BIOMED all	0.1286	0.1460	0.1065	0.0401	0.0310
	All 1	0.2475	0.2578	0.2112	0.0973	0.0611
	All 2	0.0833	0.0936	0.0475	0.0059	0.0111
	All 3	0.0412	0.0458	0.0266	0.0036	0.0059
	All	0.1310	0.1396	0.1015	0.0388	0.0279
Prec@5	BIBLE all	0.0338	0.0343	0.2174	0.0127	0.0164
	EUROPARL all	0.0436	0.0465	0.1913	0.0237	0.0150
	BIOMED all	0.0323	0.0360	0.2059	0.0215	0.0204
	All 1	0.0721	0.0767	0.3811	0.0462	0.0351
	All 2	0.0170	0.0180	0.1069	0.0041	0.0072
	All 3	0.0056	0.0079	0.0575	0.0039	0.0028
	All	0.0338	0.0365	0.1929	0.0195	0.0161

Wyniki (Potential, BOW@5)



Manchester
Metropolitan
University



Metric	Subset	Plainifier		Human	Baselines	WordNet
		Default	Tuned		SimplePPDB	
Potential	BIBLE all	0.9225	0.9366	0.2174	0.1291	0.1150
	EUROPARL all	0.8668	0.8765	0.1913	0.2228	0.0969
	BIOMED all	0.8414	0.8656	0.2059	0.1855	0.0995
	All 1	0.9885	0.9908	0.3811	0.4296	0.1755
	All 2	0.8175	0.8406	0.1069	0.0386	0.0643
	All 3	0.0412	0.0458	0.0266	0.0197	0.0310
	All	0.6394	0.6451	0.0575	0.1767	0.0952
BOW@5	BIBLE all	0.1116	0.1243	0.2085	0.0674	0.0990
	EUROPARL all	0.1051	0.1140	0.1607	0.0925	0.0750
	BIOMED all	0.0927	0.1138	0.1777	0.0815	0.0779
	All 1	0.1364	0.1369	0.1980	0.0834	0.0598
	All 2	0.0853	0.1068	0.1742	0.0785	0.0979
	All 3	0.0821	0.1095	0.2188	0.1428	0.1789
	All	0.1031	0.1187	0.1964	0.0997	0.1083

Ograniczenia



Manchester
Metropolitan
University



- Rozmiar zbioru MWLS1 (7059 podmian dla 1462 zdań) wystarczający do ewaluacji, ale zbyt mały do trenowania dużych modeli, np. Transformer,
- Jakość uproszczeń nie tak wysoka, jak w korpusach przygotowanych przez specjalistów,
- Plainifier generuje uproszczenia pasujące doskonale do kontekstu, ale często słabo zachowujące pierwotne znaczenie,
- Miary ewaluacyjne nie odzwierciedlają naszych celów: albo zbyt restrykcyjne, albo zbyt słabe (por. BLEU),
- Wciąż ogromna odległość do pokonania od umiejętności ludzi,
- Możliwość rozwiązania zadania?



Manchester
Metropolitan
University



Dziękuję!

<https://github.com/piotrrmp/mwls1>
<https://github.com/piotrrmp/plainifier>
<https://github.com/piotrrmp/tersebert>

Praca finansowana z programu *Polskie Powroty* Narodowej Agencji Wymiany Akademickiej (NAWA), grant nr PPN/PPO/2018/1/00006 oraz grantu obliczeniowego udzielonych przez *Poznańskie Centrum Superkomputerowo-Sieciowe*.

Literatura (1)



Manchester
Metropolitan
University



- Gustavo H Paetzold and Lucia Specia. Unsupervised Lexical Simplification for Non-Native Speakers. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3761–3767. AAAI Press, 2016.
- Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. Expertise Style Transfer: A New Task Towards Better Communication between Experts and Laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.100. URL <https://www.aclweb.org/anthology/2020.acl-main.100>.
- Matthew Shardlow. *Lexical simplification: optimising the pipeline*. PhD thesis, University of Manchester, UK, 2015. URL <http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.677740>.
- Zhemín Zhu, Delphine Bernhard, and Iryna Gurevych. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China, 2010. Coling 2010 Organizing Committee. URL <https://www.aclweb.org/anthology/C10-1152>.
- Matthew Shardlow. Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1583–1590, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/479_Paper.pdf.
- Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679. Association for Computational Linguistics, 2020. URL <http://arxiv.org/abs/2005.00481>.

Literatura (2)

Manchester
Metropolitan
University

- Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1331. URL <https://www.aclweb.org/anthology/P19-1331>.
- Christos Christodouloupoulos and Mark Steedman. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395, 2015. ISSN 15728412. doi: 10.1007/s10579-014-9287-y.
- Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, pages 79–86, 2005.
- Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner, K. B. Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(1), 2012. ISSN 14712105. doi: 10.1186/1471-2105-13-161.
- Sian Gooding and Ekaterina Kochmar. Complex Word Identification as a Sequence Labelling Task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/V1/P19-1109.
- Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. Lexical Simplification with Pretrained Encoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence, 2020. URL <http://arxiv.org/abs/1907.06226>.
- Ellie Pavlick and Chris Callison-Burch. Simple PPDB: A Paraphrase Database for Simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–148, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2024. URL <https://www.aclweb.org/anthology/P16-2024>.