PolEval 2021

Zadanie 4: Question Answering Challenge

Maciej Ogrodniczuk | Zespół Inżynierii Lingwistycznej
Piotr Przybyła | Instytut Podstaw Informatyki PAN

Warsztat PolEval @ AI & NLP Day

25 października 2021 r.

O co chodzi?

Czy to nowy pomysł?

Systemy QA:

- Prehistoria (1960–1990): ręcznie przygotowane reguł przekształcające zdania na zapytania do bazy danych,
- Historia (1990–2015): automatyczne zbieranie wiedzy ze stron WWW,
- Teraźniejszość (2015–2020): j.w., z głębokim uczeniem,
- Przyszłość (2021–): odpowiadanie na podstawie wiedzy zakumulowanej w pretrenowanych modelach językowych (GPT-3 etc.)

A po polsku?

Systemy QA:

- Prehistoria (1960–1990): interfejs językowy do bazy ORBIS Vetulani (1988),
- Historia (1990–2015): systemy na bazie Wikipedii (RAFAEL – Przybyła, 2016) i innych stron WWW (Walas and Jassem, 2010; Marcińczuk et al., 2013),
- Teraźniejszość (2015–2020): ??,
- Przyszłość (2021–): ??

Kolekcje pytań:

- Z panelu *Czy wiesz. . . ?* Wikipedii (Marcińczuk et al., 2013): 4721 pytań, artykuł (lub fragment) z odpowiedzią,
- Z zestawu dla potencjalnych uczestników w *Jeden z dziesięciu* (Przybyła, 2016): 1130 pytań z odpowiedziami.

Specyfika naszego zadania

Jak w teleturnieju:

- szukamy rozwiązań, które są w stanie odpowiadać na pytania z wiedzy ogólnej,
- w większości przypadków pytamy o proste fakty, tj. pojedyncze byty: osoby, miejsca, nazwy (ale niekoniecznie tylko nazwy własne),
- nie wymagamy rozumowania (*Jaki jest największy kraj z tych, które wydają więcej niż 10% PKB na zbrojenia?*),
- zadajemy pytania w języku polskim, poprawne składniowo,
- oczekujemy prostej odpowiedzi, a nie definicji czy rozbudowanego wyjaśnienia (*Czym jest globalne ocieplenie?*).

Zbieranie danych

Z wielu źródeł:

- ze stron, na których fani wymieniają się pytaniami (np. `http://zenzycia.blogspot.com/2018/08/pytania-z-teleturnieju-1-z-10-spis.html`),
- z serwisów wymiany plików (np. `chomikuj.pl`),
- z różnych quizów online (np. `onet.pl`, `radiozet.pl`, `kurierlubelski.pl`, `gazetawroclawska.pl`, `se.pl`, `polskatimes.pl`),
- z nagrań odcinków teleturniejów telewizyjnych dostępnych na platformach VOD (m.in. *To był rok*, *Va Banque* i *Jeden z dziesięciu* na `vod.tvp.pl`).
- z zasobów systemu RAFAEL.

Czyszczenie danych

Praca ręczna:

- poprawianie sformułowań,
- usuwanie duplikatów,
- weryfikacja odpowiedzi,
- dodanie wariantów odpowiedzi,
- dodanie różnych sformułowań tej samej odpowiedzi,

Zbiór danych

Przygotowaliśmy:

- zbiór rozwojowy (1000 pytań i odpowiedzi),
- zbiór testowy A (2500 pytań, bez odpowiedzi; wyniki były prezentowane na stale aktualizowanej liście chwały),
- zbiór testowy B (2500 pytań, też bez odpowiedzi, zostały użyte w ostatnim tygodniu konkursu do końcowej ewaluacji).

Nie przygotowaliśmy:

- danych treningowych!
- ale można używać wszystkich dostępnych offline zasobów (oprócz ludzkich).

Zbiór danych

Rodzaje odpowiedzi:

- konkretna wartość:
  - Q: *Jak nazywa się bohaterka gier komputerowych z serii Tomb Raider?*
  - A: *Lara Croft*
- kategoria wartości:
  - Q: *Paź królowej to gatunek których owadów?*
  - A: *motyli*
- „tak” lub „nie”:
  - Q: *Czy w przypadku skrócenia kadencji Sejmu ulega skróceniu kadencja Senatu?*
  - A: *tak*
- wartość liczbowa:
  - Q: *Ile pełnych tygodni ma rok kalendarzowy?*
  - A: *52* (nie: *pięćdziesiąt dwa*)

Zbiór danych

Rodzaje odpowiedzi:

- jedna z podanych w pytaniu wartości:
  - Q: *Co zabiera Wenus więcej czasu: obieg dookoła Słońca czy obrót dookoła osi?*
  - A: *obrót dookoła osi*
- uzupełnienie podanego zadania, cytatu, wyrażenia:
  - Q: *Proszę dokończyć powiedzenie: „piłka jest okrągła, a bramki są...”*
  - A: *dwie*
- alternatywna nazwa:
  - Q: *Jaki przydomek nosił Ludwik I, król Franków i syn Karola Wielkiego?*
  - A: *Pobożny*

Zbiór danych

Przykładowe sposoby zadawania pytań:

- *Proszę rozwinąć skrót CIA.*
- *Tę samą nazwę noszą: pocisk miotany ręką, kamień półszlachetny i owoc południowy. Jaką?*
- *Ten urodzony w XIX w. Nantes francuski pisarz uchodzi za prekursora literatury fantastycznonaukowej. O kogo chodzi?*
- *Festiwal filmowy „Dwa Brzegi" odbywa się jednocześnie w dwóch miejscowościach na dwóch brzegach Wisły. Jedna z nich to Janowiec, a druga?*
- *Tadeusz Andrzej Bonawentura, walczył o niepodległość USA. Jak brzmi jego nazwisko?*
- *Zobaczyć... i umrzeć – o które miasto chodzi?*

Zbiór danych

Jeszcze o odpowiedzi:

- powinna ograniczać się do najwyżej kilku słów
  (nie odpowiadamy całym zdaniem, linkiem, dokumentem),
- może zawierać przyimki:
  Q: *Do którego państwa należą Wyspy Eolskie?*
  A: *do Włoch*
- może być odmieniona:
  Q: *Symbolem którego pierwiastka jest Cr?*
  A: *chromu*
- może zawierać znaki interpunkcyjne:
  Q: *Która baśń Andersena opisuje dzieciństwo pewnego łabędzia?*
  A: *„Brzydkie kaczątko”*
- w przypadku pytań o osoby, zawiera pełne imię i nazwisko:
  Q: *Kto śpiewa piosenkę „Eurydyki tańczące”?*
  A: *Anna German*

Zbiór danych

Jakich pytań nie ma w zbiorze?

- o kwestie, które się szybko dezaktualizują (*Kto jest prezydentem Francji?*)
- na które odpowiedź zawiera więcej niż dwa elementy (jeśli odpowiedź zawiera dokładnie dwa elementy, należy je podać w dowolnym porządku, oddzielone spójnikiem „i”):
  Q: *Co występuje w powiedzeniu razem z makiem i oznacza nic?*
  A: *pasternak i figa / figa i pasternak*
- wymagających wybrania elementów z jakiegoś zbioru (*Proszę podać nazwy dwóch państw, przez które przepływa Nil.*)
- o pisownię (*Przez jakie ‘h’ piszemy słowo ‘charyzma’?*)
- wymagających dłuższych lub niejednoznacznych wyjaśnień (*Jakie pokrewieństwo łączyło reżysera Jana Łomnickiego i aktora Tadeusza?*)

Zbiór danych

			Dev	Test A	Test B	Suma
Pytań			1000	2500	2500	6000
Słów w pytaniu		**(średnio)**	8.43	8.41	9.49	8.87
		(mediana)	8	8	9	8
		(min.)	3	3	3	3
		(max)	21	22	32	32
Pytań z	1	**odpowiedzi**	890	2178	1890	4958
	2		104	307	570	981
	3		5	13	29	47
	4		0	1	11	12
	5		1	1	0	2

Ewaluacja

Całego systemu:

- porównujemy odpowiedzi systemu z kluczem,
- obliczamy dokładność systemu jako liczbę dobrych odpowiedzi dzieloną na rozmiar pliku testowego.

Pojedynczej odpowiedzi:

- dla pytań liczbowych (np. *W którym roku...*) będziemy obliczać zgodność wartości liczbowej wyekstrahowanej z pytania i odpowiedzi (wyrażeniem regularnym),
- dla pozostałych będziemy obliczać podobieństwo tekstowe pytania i odpowiedzi zapisanych małymi literami; do porówania będziemy używać miary Levenshteina; odpowiedź jest „zaliczona”, jeśli obliczona różnica jest mniejsza niż połowa długości którejkolwiek odpowiedzi z klucza.

Baseline `WIKI_SEARCH`

Odpytujemy Wikipedię:

1. Podziel pytanie na tokeny za pomocą spaCy (model `pl_core_news_sm`) i zignoruj tokeny jednoznakowe.
2. Wyślij tokeny oddzielone spacjami jako zapytanie do Search API polskiej Wikipedii.
3. Dla każdego ze zwróconych artykułów:
   1. podziel jego tytuł na tokeny za pomocą spaCy,
   2. jeśli żaden z tokenów tytułu nie pokrywa się w co najmniej 50% (mierzone jak podczas ewaluacji) z żadnym z tokenów zapytania:
      - jeśli w tytule znaleziono nawias otwierający, usuń część tytułu zaczynającą się od "(",
      - zwróć tytuł jako odpowiedź,
   3. w przeciwnym razie kontynuuj do następnego wyniku.
4. Jeśli jeszcze nie została znaleziona odpowiedź, usuń pierwszy token pytania i wróć do punktu 2.

Wyniki

Submission	**test-B accuracy**
Mateusz Piotrowski	71.68%
Aleksander Smywiński-Pohl et al.	50.96%
Piotr Rybak	50.96%
Darek Kłeczek	46.44%
Karol Gawron	36.12%
`WIKI_SEARCH`	9.60%
Filip Graliński	4.16%
BI Insight	0.96%

POLEVAL 2021
TASK 4 WINNER

MATEUSZ PIOTROWSKI

Shared Task 4:
Quiz Question Answering Challenge

Bibliografia

Marcińczuk M., Radziszewski A., Piasecki M., Piasecki D. and Ptak M. (2013). Evaluation of a Baseline Information Retrieval for a Polish Open-domain Question Answering System. In Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2013), pp. 428–435. Association for Computational Linguistics.

Przybyła P. (2016). Boosting Question Answering by Deep Entity Recognition. arXiv:1605.08675.

Vetulani Z. (1988). PROLOG Implementation of an Access in Polish to a Data Base. In Studia z automatyki XII, pp. 5–23. PWN.

Walas M. and Jassem K. (2010). Named Entity Recognition in a Polish Question Answering System. In Kłopotek M. A., Marciniak M., Mykowiecka A., Penczek W. and Wierzchoń S. T. (eds.), Intelligent Information Systems, pp. 181–191. Publishing House of University of Podlasie.