Odpowiadanie na pytania z użyciem automatycznie zgromadzonej bazy pojęć

Piotr Przybyła

IPI PAN

16 marca 2015

Plan prezentacji

PhD STUDIES

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Project co-financed by European Union within the framework of European Social Fund

Odpowiadanie na pytania

(ang. Question Answering, QA, Q&A)

zdolność systemu komputerowego do komunikacji z użytkownikiem w jego języku naturalnym

- prawie nieograniczone zastosowania przy zadawaniu odpowiednich pytań (*Gdzie upadnie pocisk?*, *Czy jutro będzie padać?*),
- główne problemy w obszarze **przetwarzania języka naturalnego (NLP)** i **wyszukiwania informacji tekstowych (IR & IE)**,
- poszukiwana informacja uznawana za daną w **bazie wiedzy**,
- postrzegane jako oczywista przyszłość komputerów w kulturze popularnej.

Pytania

Pytania o proste fakty (ang. **factoids**), tj. osoby, miejsca, nazwy, etc.

- bez rozumowania lub obliczeń (*Jaki jest największy kraj spośród tych, które wydają więcej niż 10% PKB na zbrojenia?*),
- poprawne składniowo pytania w języku polskim,
- odpowiedzi zawarte są (być może przeformułowane) w bazie wiedzy,
- nie oczekujemy definicji, wyjaśnień czy innych rozbudowanych odpowiedzi (*Czym jest globalne ocieplenie?*).

Niekoniecznie tylko nazwy własne (ang. named entities)!

„Jeden z dziesięciu”

- Jak nazywa się wzgórze w Paryżu na którym zbudowano bazylikę Sacré-Cœur?
- Kim zajmuje się wiktymologia?
- W którym kabarecie występuje Katarzyna Pakosińska?
- Która roślina w Inwokacji "Pana Tadeusza" Adama Mickiewicza pałała panieńskim rumieńcem?
- Jak nazywał się niemiecki teolog, który ogłosił 95 tez, krytykując sprzedawanie odpustów na rzecz budowy bazyliki Św. Piotra?
- Jak długo trwa prezydencja państwa w Unii Europejskiej?
- Co jest stolicą Tajwanu?

Przykład konfrontacji

Kiedy Albert Einstein opublikował **ogólną** teorię względności?

25 listopada 1915 r. Einstein opublikował swoją najważniejszą pracę: ogólną teorię względności.

Kiedy Albert Einstein opublikował **szczególną** teorię względności?

Rok 1905 jest określany jako Annus mirabilis (cudowny rok) Einsteina. Wtedy ten nieznany w środowisku fizyków szwajcarski urzędnik patentowy opublikował kilka prac przełomowych dla fizyki. Jego publikacja „Zur Elektrodynamik bewegter Körper" („O elektrodynamice ciał w ruchu") wprowadza nową teorię, nazwaną później szczególną teorią względności, która ...

Odpowiedzi

Cztery poziomy udzielania odpowiedzi:

1. Wskazanie dokumentu zawierającego odpowiedź,
   (*Wikipedia:Albert_Einstein*)
2. Fragment tekstu,
   (jedno lub więcej zdań, jak w przykładzie)
3. Tylko jedna nazwa,
   (*25 listopada 1915 r.*)
4. Pełne zdanie. (*Albert Einstein opublikował ogólną teorię względności 25 listopada 1915 r.*)

Information Retrieval (**IR**) vs. Information Extraction (**IE**).

Odpowiadanie na poziomie nazw

W konsekwencji wyższe wymagania względem:

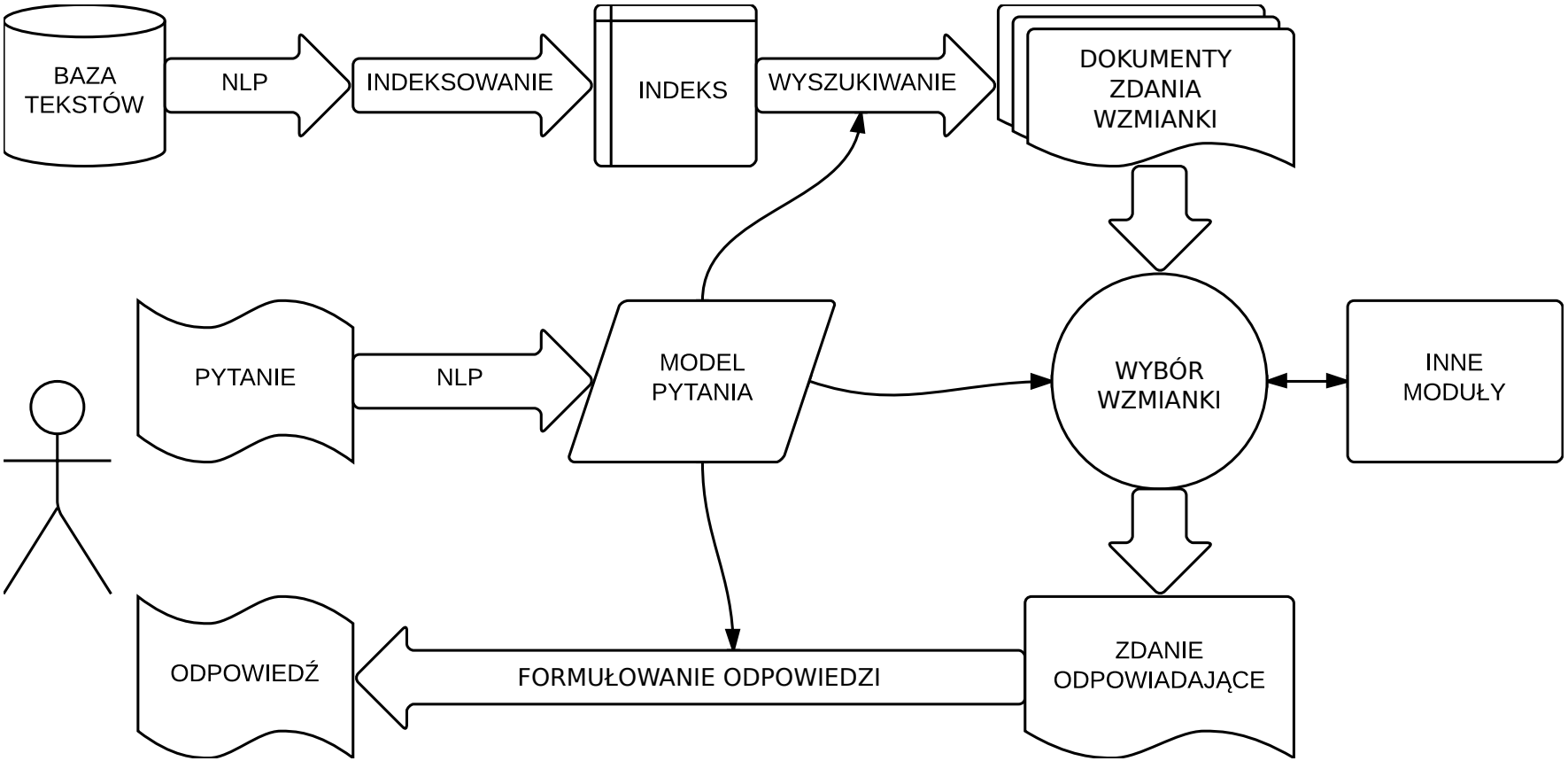
- analizy pytania,
  (precyzyjne określenie typu poszukiwanej nazwy)
- wyboru odpowiedzi,
  (wyodrębnienie nazwy zadanego typu w tekście źródłowym)
- ewaluacji.
  (weryfikacja zgodności odpowiedzi uwzględniająca wielość sformułowań)

$\Longrightarrow$ główne ograniczenia projektu RAFAELa.

RApid Factoidal Answer Extracting aLgorithm

- system odpowiadania na pytania bez ograniczeń dziedziny na podstawie tekstowej bazy wiedzy,
- opiera się na bazie narzędzi językowych opracowanych wokół NKJP (i nie tylko),
- budowany dla celów badawczych – wielowariantowy,
- udziela odpowiedzi, podając nazwę poszukiwanego bytu.
- [Przybyła, 2013b, Przybyła, 2012, Przybyła, 2013a]

Uproszczony schemat architektury RAFAELa

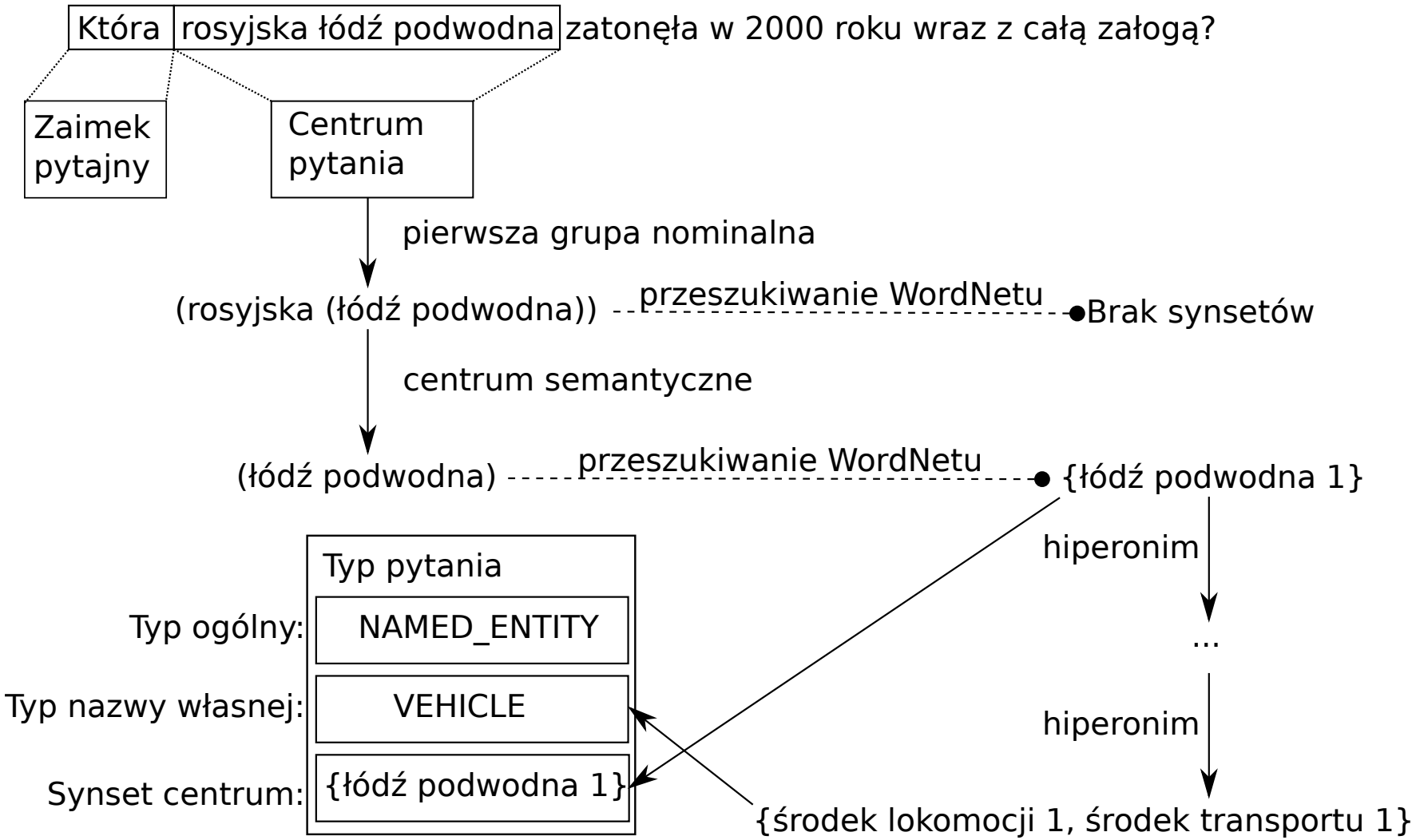


Analiza pytania

Służy przetworzeniu zdania pytającego na **model pytania** (*Stolicą którego państwa jest Chartum?*):

- **Typ pytania** – jaka informacja poszukiwana?
  - **Ogólny typ** – ze względu na sposób szukania odpowiedzi `NAMED_ENTITY`.
  - **Typ nazwy własnej** – do narzędzi rozpoznających nazwy własne `CITY`.
  - **Synset centrum** – do głębokiego rozpoznawania nazw <*stolica.1, miasto stołeczne.1*>.
- **Zapytanie** – do wyszukania dokumentów zgodnych z tematem `(stolica państwo jest Chartum)`$\sim$`1`.
- **Treść pytania** – słowa oczekiwane w kontekście odpowiedzi {`stolica, jest, Chartum`}.

Określenie typu pytania



Budowa zapytania do wyszukiwania

1. Usunięcie elementów konstrukcyjnych (*którego, ?*),
2. Zachowanie centrum pytania (*państwa*),
3. Połączenie operatorem alternatywy (*stolica państwo jest Chartum*)$\sim$1.
4. Rozmyte dopasowywanie – *fuzzy term matching* w *Lucene*: dopasowanie, gdy dwa słowa w odległości edycyjnej Levenshteina co najwyżej 3, przy czym pierwsze (n-3) znaków zgodne.

Zapytanie skierowane do indeksu *Lucene*, z listy rankingowej wybieramy pierwszych N dokumentów do dalszej analizy.

Rozpoznawanie nazw

1. Bazujące na wcześniejszym znakowaniu narzędziami NER:
   1.1 *Nerf* 0.1 – 13 typów nazw (z zagnieżdżeniami)
   1.2 *Liner2* 2.3 – 56 typów nazw (można zredukować do 5 dla lepszej precyzji)
2. Własne narzędzie do rozpoznawania liczb i wielkości (*Quant*),
3. Głębokie rozpoznawanie nazw (DeepER),
4. Rozwiązanie hybrydowe, łączące wzmianki ze wszystkich źródeł.

Mapowanie typów (fragment)

Typ NE pytania	Typ Nerf	Typ Liner2	Typ Quant
CONTINENT	geogName	continent_nam	
LAKE			
MOUNTAIN			
ARCHIPELAGO			
SEA		sea_nam	
PERSON	persName	person_nam	
NAME	persName:forename		
SURNAME	persName:surname		
BAND	orgName	band_nam	
DYNASTY	persName:addName		
ORGANISATION	orgName	organization_nam institution_nam political_party_nam	
EVENT		event_nam	
TIME	date		
CENTURY			
YEAR			
COUNT			number
QUANTITY			quantity
VEHICLE			

Wybór wzmianki

1. Wybór kontekstu:
   1.1 zdanie zawierające wzmiankę,
   1.2 usunięta treść wzmianki,
   1.3 dodany tytuł dokumentu (anafory).
2. Pomiar podobieństwa **treści** pytania i **kontekstu** odpowiedzi.
3. Wybór wzmianki o najwyższej wartości.

Indeks podobieństwa Jaccarda:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|},$$

Ważony IDF-ami słów:

$$\mathrm{Sim}_w(A,B) = \frac{\sum_{i \in A \cap B} w_i}{\sum_{i \in A \cup B} w_i}, \quad w_i = \frac{\log \frac{|D|}{|\{d : i \in d\}|}}{\max_j \log \frac{|D|}{|\{d : j \in d\}|}},$$

DeepER vs NER

Głębokie rozpoznawanie nazw (ang. *Deep Entity Recognition*, DeepER) – uogólnienie rozpoznawania nazw własnych przez zastąpienie kategorii przez precyzyjne synsety WordNet.

- bardziej precyzyjna selekcja wzmianek:
  Który europejski monarcha powrócił do kraju jako premier republiki?
  Symeon II $\Longrightarrow$ <car.1> $\Longrightarrow$ <monarcha.1, koronowana głowa.1>
  Nie: `NAMED_ENTITY:PERSON`
- odpowiadanie na pytania spoza kategorii nazw własnych:
  32% w JzD: zwierzęta, rasy, pojęcia abstrakcyjne, rośliny, substancje chemiczne, zwyczaje, procesy, tkanki, procesy historyczne, urządzenia, zjawiska społeczne, przedsiębiorstwa, ...

Biblioteka nazw

Gromadzi deskryptory bytów w następującym formacie (wpis #9751):

- główna nazwa: *Bronisław Komorowski*,
- inne nazwy (aliasy): *Bronisław Maria Komorowski*, *Komorowski*,
- synsety WordNet:
  - <podsekretarz.1, podsekretarz stanu.1, wiceminister.1>,
  - <wicemarszałek.1>,
  - <polityk.1>,
  - <wysłannik.1, poseł.1, posłaniec.2, wysłaniec.1, posłannik.1>,
  - <marszałek.1>,
  - <historyk.1>,
  - <minister.1>,
  - <prezydent.1, prezydent miasta.1>.

Dla potrzeb RAFAELa stworzono bibliotekę składającą się z 809.786 bytów, opisanych przez 1.169.452 nazwy (975.592 różne) i 1.264.918 synsetów (31.545 różnych).

Budowa biblioteki

Pozyskiwanie deskryptorów z encyklopedii, korzystając z:

- definicji:
  Lech Wałęsa (ur. 29 września 1943 w Popowie) – polski polityk i działacz związkowy. Współzałożyciel i pierwszy przewodniczący „Solidarności", opozycjonista w okresie PRL. Prezydent Rzeczypospolitej Polskiej w latach 1990–1995, laureat Pokojowej Nagrody Nobla (1983)[1], przez tygodnik „Time" uznany za Człowieka Roku (1981) oraz za jednego ze 100 najważniejszych ludzi stulecia (1999)
- stron przekierowań (dodatkowe nazwy),
- stron ujednoznaczniających (dodatkowe nazwy i definicje):
  Wałęsa: Lech Wałęsa (ur. 1943) – przywódca Solidarności, Prezydent RP w latach 1990–1995, laureat Pokojowej Nagrody Nobla

Przykład

Pierwszy akapit artykułu:

> *Lech Wałęsa (ur. 29 września 1943 w Popowie) – polski polityk i działacz związkowy. Współzałożyciel i pierwszy przewodniczący „Solidarności", opozycjonista w okresie PRL. Prezydent Rzeczypospolitej Polskiej w latach 1990–1995, laureat Pokojowej Nagrody Nobla (1983)[1], przez tygodnik „Time" uznany za Człowieka Roku (1981) oraz za jednego ze 100 najważniejszych ludzi stulecia (1999).*

Przykład

Krok 1: usunięcie tekstu w nawiasach i cudzysłowach:

> *Lech Wałęsa ~~(ur. 29 września 1943 w Popowie)~~ – polski polityk i działacz związkowy. Współzałożyciel i pierwszy przewodniczący ~~„Solidarności”~~, opozycjonista w okresie PRL. Prezydent Rzeczypospolitej Polskiej w latach 1990–1995, laureat Pokojowej Nagrody Nobla ~~(1983)[1]~~, przez tygodnik ~~„Time”~~ uznany za Człowieka Roku ~~(1981)~~ oraz za jednego ze 100 najważniejszych ludzi stulecia ~~(1999)~~.*

Przykład

Krok 2: wydobycie definicji na podstawie wzorca:

> ***(****Lech Wałęsa****)*** **–** ***(****polski polityk i działacz związkowy. Współzałożyciel i pierwszy przewodniczący, opozycjonista w okresie PRL. Prezydent Rzeczypospolitej Polskiej w latach 1990–1995, laureat Pokojowej Nagrody Nobla, przez tygodnik uznany za Człowieka Roku oraz za jednego ze 100 najważniejszych ludzi stulecia.****)***

Oprócz myślników usuwane również wyrażenia takie, jak: *jest to*, *to*, *jeden z*, *gatunek*, etc.

Przykład

Krok 3: podzielenie według separatorów:

polski polityk i działacz związkowy .
Współzałożyciel i pierwszy przewodniczący ,
opozycjonista w okresie PRL .
Prezydent Rzeczypospolitej Polskiej w latach 1990–1995 ,
laureat Pokojowej Nagrody Nobla ,
przez tygodnik uznany za Człowieka Roku oraz za jednego ze . . . .

Separatory: średnik, przecinek i kropka (zdania!).

Przykład

Krok 4: redukcja każdego fragmentu do grupy nominalnej:

polski polityk i działacz związkowy

Współzałożyciel i pierwszy przewodniczący

opozycjonista ~~w okresie PRL~~

Prezydent Rzeczypospolitej Polskiej ~~w latach 1990–1995~~

laureat Pokojowej Nagrody Nobla

~~przez tygodnik uznany za Człowieka Roku oraz za jednego ze . . .~~

Jeśli któryś z fragmentów nie zaczyna się grupą nominalną, to zostaje usunięty z analizy, wraz ze wszystkimi następnikami.

Przykład

Krok 5: rozbicie grup koordynacyjnych:

polski polityk **i** *działacz związkowy*

Współzałożyciel **i** *pierwszy przewodniczący*

opozycjonista

Prezydent Rzeczypospolitej Polskiej

laureat Pokojowej Nagrody Nobla

Przykład

Krok 6: poszukiwanie odpowiedników w WordNecie:

polski ***polityk*** $\Longrightarrow$ *<polityk.1>*

działacz *związkowy* $\Longrightarrow$ *<działacz.1, aktywista.1>*

Współzałożyciel $\Longrightarrow$ *<współzałożyciel>*

pierwszy ***przewodniczący*** $\Longrightarrow$ *<przewodniczący.1>*

opozycjonista $\Longrightarrow$ *<opozycjonista.1>*

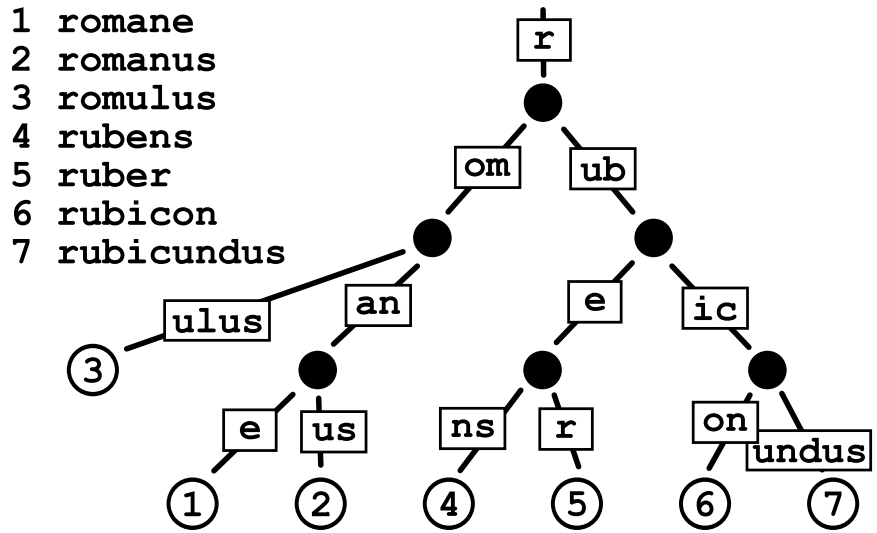
Prezydent *Rzeczypospolitej Polskiej* $\Longrightarrow$ *<prezydent.1, prezydent miasta.1>*

laureat *Pokojowej Nagrody Nobla* $\Longrightarrow$ *<laureat.1>*

Wykorzystanie głowy semantycznej każdej grupy.

Rozpoznawanie

Poszukiwanie na żądanie wystąpień każdej z 1.169.452 nazw w badanym tekście, drzewo prefiksowe PATRICIA:



Źródło: Wikipedia/Radix tree

Dla odnalezionych nazw sprawdzenie zgodności z **synsetem centrum** – czy zachodzi (pośrednia) hiperonimia?

Rozpoznawanie

Problem z nazwami własnymi: często nierozpoznana struktura składniowa, jeszcze częściej błędne lematy. Dlatego trzy źródła nazw:

1. lematy grup składniowych i słów,
2. lematy pojedynczych segmentów,
3. formy ortograficzne słów.

Dopasowanie dwóch łańcuchów zachodzi, gdy:

- wspólny prefiks,
- niedopasowane sufiksy nie dłuższe niż 3 znaki,
- wspólny prefiks stanowi większość każdego ze słów.

Automatyczna ewaluacja

Problem: niezgodność odpowiedzi oczekiwanej z otrzymaną.
Przykład: *Kto jest obecnie prezydentem Polski?*
oczekiwano *Bronisław Komorowski*, otrzymano *Komorowski*.
Rozwiązanie: głębokie rozpoznawanie nazw:

1. Pobranie listy bytów przypisanej do nazwy oczekiwanej (trzech BK: pisarz, ksiądz i polityk).
2. Pobranie wyniku rozpoznawania nazw na wybranej wzmiance (25 mężczyzn, osiem kobiet i dwie wsie).
3. Wybranie przecięcia zbiorów (dwóch BK: ksiądz i polityk).

Niepuste przecięcie $\Longrightarrow$ odpowiedź zweryfikowana (nie ujednoznaczniona!).

Ewaluacja

Dane:

1. korpus źródłowy: zawartość Wikipedii z 3 marca 2013, ∼900.000 artykułów,
2. pytania zbioru rozwojowego: 1130 pytań z teleturnieju *Jeden z dziesięciu* [Karzewski, 1997].
3. pytania zbioru testowego: 576 pytań zespołu z PWr, na podstawie rubryki *Czy wiesz . . . ?* Wikipedii [Marcińczuk et al., 2013].
4. odpowiedzi opracowane ręcznie (+ artykuł źródłowy i typ pytania).

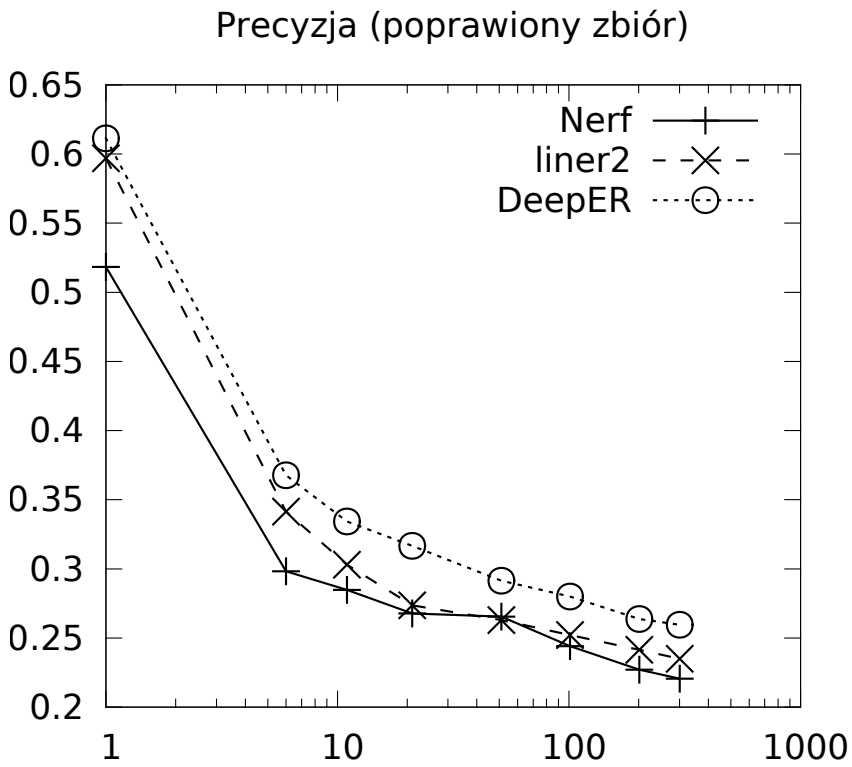
Miary oceny:

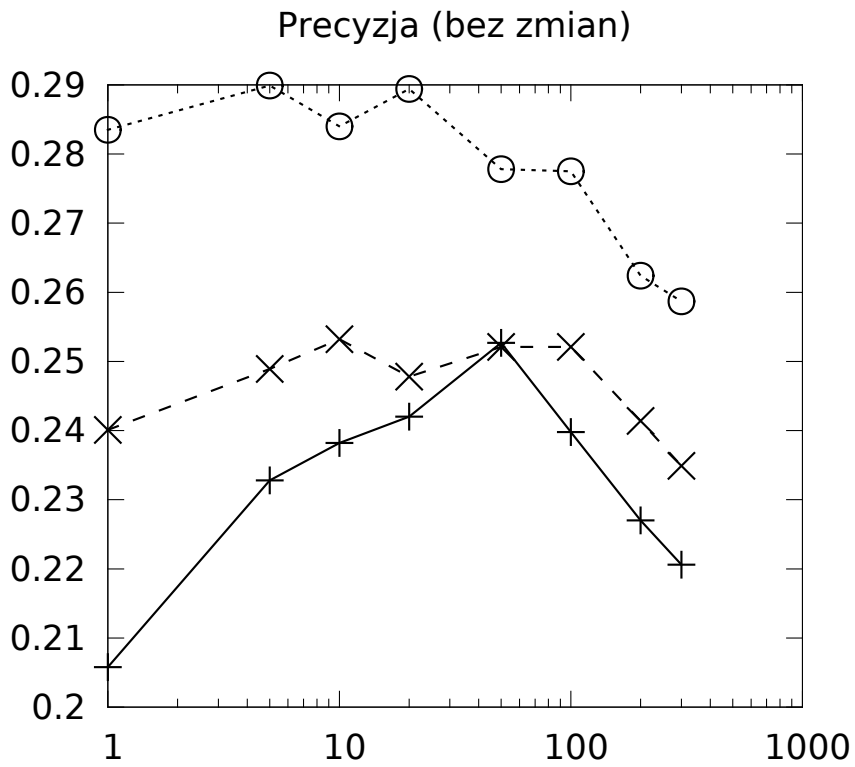
1. pokrycie: udział pytań, na które udzielono jakiejkolwiek odpowiedzi,
2. precyzja: udział poprawnych odpowiedzi,
3. miara MRR: suma odwrotności pozycji prawidłowej odpowiedzi na liście rankingowej.

Przykładowe pytania z metadanymi

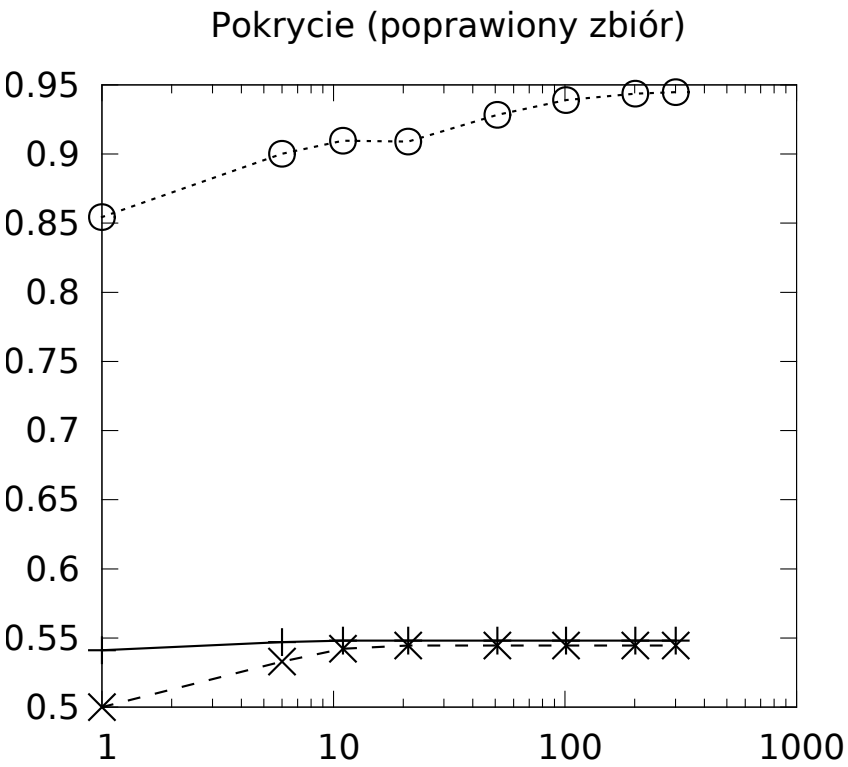
Pytanie		
Typ pytania	**Artykuł źródłowy**	**Odpowiedź**
W którym roku umarł Stefan Żeromski?		
NAMED_ENTITY:YEAR	Stefan Żeromski	1925
Jakie organella nadają barwę korzeniom marchwi?		
UNNAMED_ENTITY	Chromoplast	Chromoplasty
Jakiego wyznania jest większość mieszkańców Liechtensteinu?		
UNNAMED_ENTITY	Liechtenstein	Katolicyzm
Który z filozofów był twórcą „atomizmu”?		
NAMED_ENTITY:PERSON	Demokryt	Demokryt z Abdery
Czy Jacques Brel pochodził z Francji?		
TRUEORFALSE	Jacques Brel	Nie

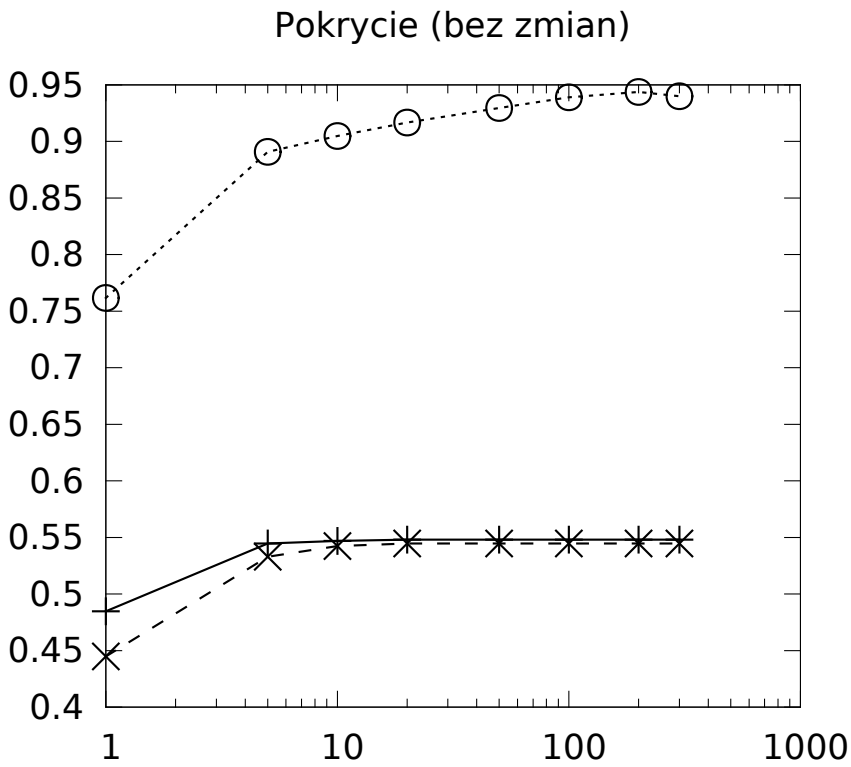
Precyzja


Precyzja (poprawiony zbiór)
Nerf
liner2
DeepER
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
1
10
100
1000


Precyzja (bez zmian)
0.29
0.28
0.27
0.26
0.25
0.24
0.23
0.22
0.21
0.2
1
10
100
1000

Pokrycie


Pokrycie (poprawiony zbiór)
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
1
10
100
1000


Pokrycie (bez zmian)
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
1
10
100
1000

Wyniki końcowe

- oddzielny zbiór pytań, nieużywany do tej pory,
- brak korekt, 20 dokumentów,
- ręczne sprawdzenie poprawności (+1-5% precyzji).

	Pokrycie	**Precyzja**	**Miara F1**	**MRR**
Nerf	56.25% ± 2.12%	34.88% ± 2.73%	0.4306 ± 0.0213	33.66% ± 2.29%
Liner2	45.31% ± 2.05%	**39.08%** ± 2.90%	0.4197 ± 0.0188	41.36% ± 2.70%
DeepER	72.92% ± 1.88%	35.24% ± 2.23%	0.4751 ± 0.0214	32.80% ± 1.99%
Hybrid	**89.58%** ± 1.24%	33.14% ± 2.01%	0.4838 ± 0.0221	35.57% ± 1.88%

Odchylenie standardowe – 500 prób zbioru testowego ze zwracaniem, obliczenie wariancji otrzymanych 500 wskaźników wydajności.

Wnioski

- Zgodnie z oczekiwaniami, technika DeepER znacząco polepsza pokrycie zbioru pytań,
  $\Rightarrow$ pytania poza nazwami własnymi,
- Różnica w precyzji ze zbioru rozwojowego nie utrzymuje się na zbiorze testowym,
  $\Rightarrow$ mniej pytań zawierających centrum,
- *Liner2* zauważalnie lepiej wpływa na precyzję niż *Nerf*,
  $\Rightarrow$ bardziej szczegółowa kategoryzacja,
- Metoda hybrydowa obsługuje 90% pytań ze zbioru.
  $\Rightarrow$ pozostają pytania wymagające innych technik: `VERIFICATION`, `WHICH`, `OTHER_NAME`, `MULTIPLE`.

Pytanie o nazwy ogólne

Pytanie	
Centrum	**Odpowiedź**
Który żleb uważany jest za najbardziej lawiniasty w całych Karpatach?	
żleb	Pusty Żleb
Jaki owad pożera liście owadożernej rosiczki?	
owad	Piórolotek bagniczek
Jaki przyrząd pozwala na pomiar współczynnika załamania światła?	
przyrząd	Refraktometr Abbego
Jaki związek chemiczny służy do otrzymywania boru o wysokiej czystości?	
związek chemiczny	Jodek boru
Jaką metodą łamano szyfry Enigmy przed wynalezieniem cyklometru?	
metoda	Metoda rusztu

Przyczyny błędów

- DeepER
  - Brak ujednoznaczniania słów:
    - słowa w treści definicji i pytań,
    - nazwy, np. *kot*: zwierzę, teatr, jezioro, wieś, odznaka, 10 osób, etc.
    - nazwy pozorne, np. nordycki bóg *Od*,
- NER
  - *Quant*: ignorowanie typów wielkości,
  - narzędzi NER: niedostateczna precyzja (małe dane treningowe),
- inne
  - Rozbicie zwięzłej treści pytania na wiele zdań w artykule,
  - Odmiana nazw własnych (!!!),
  - Brak możliwości wyboru jednej nazwy w modelu *bag of words*.

Podsumowanie

- Powstał RAFEL – pierwszy polski system QA udzielający odpowiedzi w formie nazw poszukiwanych bytów na podstawie bazy tekstowej.
- Wiele problemów wynika ze specyfiki języka: część ma proste rozwiązania (zapytania rozmyte przy wyszukiwaniu), część wymaga więcej pracy (dopasowywanie zdań).
- Uogólnienie modułu rozpoznawania nazw pozwoliło na polepszenie precyzji odpowiedzi i znaczące powiększenie zbioru obsługiwanych pytań,
- Opracowano metodę automatycznego generowania biblioteki bytów, kojarzącej nazwy z synsetami WordNet, na podstawie encyklopedii.

Efekty uboczne

Zasoby powstałe „przy okazji” projektu, ale posiadające szersze zastosowanie:

- korpus wikipedii:
  - 895.486 dokumentów, 169 milionów segmentów, 1,06 GB tekstu,
  - znakowany automatycznie na poziomie segmentacji, tagowania, parsowania powierzchniowego i rozpoznawania nazw własnych (*Nerf*+*Liner2*), 22,9 GB (skompresowane),
- zestaw 1706 pytań o proste fakty z odpowiedziami ręcznie opracowanym na podstawie korpusu źródłowego,
- biblioteka bytów, zawierająca 809.786 pojęć, kojarzących nazwy i synsety WordNet,
- usprawniona wersja gramatyki lematyzującej grupy nominalne [Degórski, 2012].

Degórski, L. (2012).

Towards the Lemmatisation of Polish Nominal Syntactic Groups Using a Shallow Grammar.
In *Proceedings of the International Joint Conference on Security and Intelligent Information Systems*, volume 7053 of *Lecture Notes in Computer Science*, pages 370–378. Springer-Verlag.

Karzewski, M. (1997).

Jeden z dziesięciu - pytania i odpowiedzi.
Muza SA.

Marcińczuk, M., Ptak, M., Radziszewski, A., and Piasecki, M. (2013).

Open dataset for development of Polish Question Answering systems.
In *Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Wydawnictwo Poznańskie, Fundacja Uniwersytetu im. Adama Mickiewicza.

Przybyła, P. (2012).

Issues of Polish Question Answering.
In Hryniewicz, O., Mielniczuk, J., Penczek, W., and Waniewski, J., editors, *Proceedings of the first conference 'Information Technologies: Research and their Interdisciplinary Applications' (ITRIA 2012)*, pages 122–139. Institute of Computer Science, Polish Academy of Sciences.

Przybyła, P. (2013a).

Question Analysis for Polish Question Answering.
In *51st Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop*, pages 96–102, Sofia, Bulgaria. Association for Computational Linguistics.

Przybyła, P. (2013b).

Question Classification for Polish Question Answering.
In Kłopotek, M. A., Koronacki, J., Marciniak, M., Mykowiecka, A., and Wierzchoń, S. T., editors, *Proceedings of the 20th International Conference on Language Processing and Intelligent Information Systems (LP&IIS 2013)*, pages 50–56. Springer-Verlag.

?