

Korpus sformalizowanych kroków wynikania tekstowego

Adam Przepiórkowski, Jakub Kozakoszczak,
Tadeusz Teleżyński, Jan Winkowski, Daniel Ziembicki

IPI PAN

24 Października, 2016

Spis treści

Wstęp

Derywacje

Reguły

Wnioski

Wstęp

Zadanie 4. projektu wspierającego CLARIN

Metodologia wspomaganie i ewaluacji systemów wykrywających relację wynikania tekstowego w języku polskim.

- ▶ Czas
- ▶ Budżet
- ▶ Stażyści

Wynikanie tekstowe

Wynikanie tekstowe (Dagan et al., 2006)

Wynikanie tekstowe zdefiniowane jest jako kierunkowa relacja między parami tekstów, oznaczanymi przez T – Tekst będący przesłanką, i H – wynikająca Hipoteza. Mówimy, że H wynika z T jeżeli człowiek czytający T przeważnie powiedziałby, że H jest najpewniej prawdziwe.

Miękkie podejście do wynikania:

- ▶ zsubiektywizowane (*człowiek czytający [...] powiedziałby*)
- ▶ statystyczne (*najpewniej prawdziwe*)

Wynikanie tekstowe (Dagan et al., 2006)

Wynikanie tekstowe zdefiniowane jest jako kierunkowa relacja między parami tekstów, oznaczanymi przez T – Tekst będący przesłanką, i H – wynikająca Hipoteza. Mówimy, że H wynika z T jeżeli człowiek czytający T przeważnie powiedziałby, że H jest najpewniej prawdziwe.

Miękkie podejście do wynikania:

- ▶ zsubiektywizowane (*człowiek czytający [...] powiedziałby*)
- ▶ statystyczne (*najpewniej prawdziwe*)

Zadanie RTE

Zadanie RTE (Recognizing Textual Entailment) przebiega następująco:

1. Pary (T, H) oznaczone wynikaniem zostają podzielone na dwa korpusy (treningowy i testowy)
2. Korpus treningowy jest publicznie udostępniany i wykorzystywany przez zespoły budujące systemy RTE.
3. Stworzone systemy zostają użyte do otagowania wynikaniem korpusu testowego.
4. Wynikiem systemu jest procent poprawnie oznaczonych przez niego par.

RTE 3 (2007)

- ▶ Question Answering (71%)
- ▶ Information Retrieval (66%)
- ▶ Text Summarization (58%)
- ▶ Information Extraction (52%)

FraCaS (1996)

- ▶ Generalized Quantifiers
- ▶ Plurals
- ▶ (Nominal) Anaphora
- ▶ Ellipsis
- ▶ Adjectives
- ▶ Comparatives
- ▶ Temporal Reference
- ▶ Verbs
- ▶ Attitudes

Przykłady par

RTE3

T: A former employee of the company, David Vance of South Portland, said Hooper spent a lot of time on the road, often meeting with customers between Portland and Kittery.

H: David Vance lives in South Portland.

FraCaS

P1: A Swede won a Nobel prize.

P2: Every Swede is a Scandinavian.

H: A Scandinavian won a Nobel prize.

Cel Projektu

Jakościowa ewaluacja systemów RTE

Zarys metodologii

- ▶ Derywacja tekstowa złożona z kroków wynikania.
- ▶ Sformalizowane reguły opisujące kroki derywacji.

Stworzone zasoby

Korpusy

Dwa korpusy derywacji tekstowych:
RTE3 (polski) i FraCaS (angielski)
Korpus reguł (PL/EN)

Korpus par	Korpus derywacji
Tekst Hipoteza	Tekst Krok atomowy Krok atomowy Krok atomowy ... Hipoteza

Derywacje

Krok atomowy

Krok atomowy to dodanie, zastąpienie lub usunięcie frazy lub zdania. Ogólne warunki nakładane na krok atomowy:

- ▶ Jak najmniejsza zmiana przybliżająca do hipotezy
- ▶ Przejście od poprawnego zdania do poprawnego zdania

Kroki atomowe. Przykłady

Usunięcie zdania

I: Bountiful przybył po zakończeniu wojny, wpływający do zatoki San Francisco 21 sierpnia 1945 roku. **Później Bountiful został przeznaczony na okręt szpitalny w Yokosuka w Japonii.** Wypłynął z San Francisco 1 listopada 1945 roku.

O: Bountiful przybył po zakończeniu wojny, wpływający do zatoki San Francisco 21 sierpnia 1945 roku. Wypłynął z San Francisco 1 listopada 1945 roku.

Kroki atomowe. Przykłady

Koreferencja

I: Komarzyce ulegają zakażeniu pasożytem malarii, kiedy piją krew osób zakażonych malarią. **Owady** mogą następnie przenosić chorobę na innych ludzi, których kęsają, przy czym same nie chorują.

O: Komarzyce ulegają zakażeniu pasożytem malarii, kiedy piją krew osób zakażonych malarią. **Komarzyce** mogą następnie przenosić chorobę na innych ludzi, których kęsają, przy czym same nie chorują.

Kroki atomowe. Przykłady

Hiperonimia

I: Oh Eun-sun była **koreanką**, która samotnie weszła na Mount Everest.

O: Oh Eun-sun była **kobietą**, która samotnie weszła na Mount Everest.

Pełne derywacje. Przykłady

Za pomocą kroków atomowych można przejść od T do H.

RTE3_d_92:

Kinston Indians to drugoligowa drużyna baseballowa z Kinston w Karolinie Północnej. Drużyna ta, stowarzyszona jako drużyna klasy A z drużyną Cleveland Indians, bierze udział w rozgrywkach Carolina League.

RTE3_d_92.1: *usuwanie przydawki*

Kinston Indians to drugoligowa drużyna baseballowa z Kinston w Karolinie Północnej. Drużyna ta bierze udział w rozgrywkach Carolina League.

RTE3_d_92.2: *koreferencja*

Kinston Indians to drugoligowa drużyna baseballowa z Kinston w Karolinie Północnej. Kinston Indians biorą udział w rozgrywkach Carolina League.

Pełne derywacje. Przykłady

RTE3_d_92:

:

Kinston Indians to drugoligowa drużyna baseballowa z Kinston w Karolinie Północnej. Kinston Indians biorą udział w rozgrywkach Carolina League.

RTE3_d_92.3: *usunięcie zdania*

Kinston Indians biorą udział w rozgrywkach Carolina League.

RTE3_d_92.4: *brać udział w X-e → być uczestnikiem X*

Kinston Indians są uczestnikami rozgrywek Carolina League.

RTE3_d_92.5: *wynikanie leksykalne (rozgrywki X-a → X)*

Kinston Indians są uczestnikami Carolina League.

Przykłady derywacji

Niektóre pary (T, H) wymagają użycia niewielu kroków atomowych.

RTE3_d_481:

Gurkowie pochodzą z górzystego Nepalu i są wojownikami znanymi z wyjątkowej nieustępliwości - jako piechurzy są najlepsi w walce bezpośredniej prowadzonej w dżungli/górach oraz w walce wręcz.

RTE3_d_481.1: *redukcja do członu koniunkcji*

Gurkowie pochodzą z górzystego Nepalu.

RTE3_d_481.2: *usunięcie przydawki*

Gurkowie pochodzą z Nepalu

Przykłady derywacji

FraCaS obrazuje podręcznikowe przykłady logicznych zjawisk, których przełożenie na kroki atomowe bywa zaskakująco długie.

FCS_018:

Every European has the right to live in Europe. Every European is a person. Every person who has the right to live in Europe can travel freely within Europe.

FCS_018.1: R013.1

Every European is a European who has the right to live in Europe. Every European is a person. Every person who has the right to live in Europe can travel freely within Europe.

FCS_018.2: R202

Every European is a European who has the right to live in Europe. Every European who has the right to live in Europe is a person who has the right to live in Europe. Every person who has the right to live in Europe can travel freely within Europe.

FCS_018.3: R26.3: syllogism AB.BC.AC

Every European is a person who has the right to live in Europe.

Przykłady derywacji

FCS_008:

Many great tenors are German.

FCS_008.1: *R010*

There are great tenors who are German.

FCS_009:

Several great tenors are British.

FCS_009.1: *R010*

There are great tenors who are British.

FCS_010:

Most great tenors are Italian.

FCS_010.1: *R010*

There are great tenors who are Italian

Procedura anotacji

RTE3:

- ▶ Dwie osoby opracowują derywację niezależnie od siebie.
- ▶ Opracowują wspólną wersję.
- ▶ Jeżeli stażyści opracowywali derywacje, sprawdzała ją trzecia osoba (nie-stażysta).

FraCaS:

- ▶ Derywację opracowuje jedna osoba.
- ▶ Druga osoba sprawdza.
- ▶ Stażyści nie brali udziału.

Reguły

Reguła wynikania tekstowego

Reguła wynikania tekstowego **sankcjonuje składniowo** relację wynikania tekstowego między dwoma zdaniami lub dłuższymi tekstami. Wyjście wynika z wejścia, jeśli pasują do **wzorców składniowych** zdefiniowanych przez regułę. Reguła może też wymagać własności morfosyntaktycznych, leksykalnych, semantycznych, logicznych, kotekstowych i kontekstowych.

Przykład 1

I: Nawet 100 milionów dzieci zatrudnia się w Azji.

O: Nawet 100 milionów nieletnich zatrudnia się w Azji.

R251.6: Zastęp hiperonimem

L: en pl

Formalizacja:

LHS:

X1+? [1]

RHS:

Y1 [1]

C:

hypernym(Y1, X1)

UE({X1})

T: Sub

S: S

Opis formalizmu

Reguła to **czwórka uporządkowana**: nazwa, komentarz, język i formalizacja relacji między dwoma tekstami.

R251.6: Zastęp hiperonimem

L: en pl

Formalizacja:

LHS:

X1+? [1]

RHS:

Y1 [1]

C:

hypernym(Y1,X1)

UE({X1})

T: Sub

S: S

Opis formalizmu

Reguła to czwórka uporządkowana: nazwa, komentarz, język i **formalizacja** relacji między dwoma tekstami.

Formalizacja to **piątka uporządkowana**: LHS, RHS, ograniczenia pozaskładniowe, typ (substytutywność) i siła wynikania.

Usuń zdanie (en pl)

LHS:

X1

-root-> X2?+

RHS:

X1

C:

T: Sub

S: S

Opis formalizmu

Formalizacja to piątka uporządkowana: **LHS**, **RHS**, ograniczenia pozaskładniowe, typ (substytucyjność) i siła wynikania.

Wzorzec składniowy to zbiór częściowych drzew zależnościowych: dla wejścia (LHS) i dla wyjścia (RHS).

Usuń zdanie (en pl)

LHS:

X1

-root-> X2?+

RHS:

X1

C:

T: Sub

S: S

Opis formalizmu

Wzorzec składniowy to zbiór częściowych drzew **zależnościowych**: dla wejścia (LHS) i dla wyjścia (RHS).

Typy zależności pochodzą z **Universal Dependencies**.

Usuń zdanie (en pl)

LHS:

X1

-root-> X2?+

RHS:

X1

C:

T: Sub

S: S

Opis formalizmu

Formalizacja to piątka uporządkowana: LHS, RHS, **ograniczenia pozaskładniowe**, typ (substytucyjność) i siła wynikania. Ograniczenia wyrażone są w logice pierwszego rzędu z identycznością.

I: Śmierć uwolniła Swifta z apatii w 1745 r.

O: Swift zmarł w 1745 r.

C:

UE({X1})

tense(Y1) = tense(X1)

lemma(X1) = UWOLNIĆ; lemma(X2) = ŚMIERĆ; lemma(X5) =

Z; lemma(Y1) = ZEMRZEĆ

Opis formalizmu

Formalizacja to piątka uporządkowana: LHS, RHS, ograniczenia pozaskładniowe, **typ (substytucyjność)** i siła wynikania.

syllogism AB.BC.AC for proper name A (en)

I: Mary is a student. Every student used her workstation.

O: Mary is a student. Every student used her workstation. Mary used her workstation.

Opis formalizmu

syllogism AB.BC.AC for proper name A (en)

LHS:

X1 [1]

-cop-> X2

-nsubj-> X3 [2]

X5 [3]

-nsubj-> X6 [1]

-det-> X7

RHS:

X5 [3]

-nsubj-> X3 [2]

T: Add

Opis formalizmu

O: Mary is a student. **Every** student used her workstation. Mary used her workstation.

syllogism AB.BC.AC for proper name A (en)

LHS:

X1 [1]

-cop-> X2

-nsubj-> X3 [2]

X5 [3]

-nsubj-> X6 [1]

-det-> **X7**

[...]

C: lemma(**X7**) in {EVERY,EACH,ALL,ANY}

Przykład: diateza

I: Software faults were blamed for the system failure.

O: The system failure was blamed on software faults.

diathesis for 'blame' (en)

LHS:

X1 [1]
-nsubjpass-> X2 [2]
-iobj-> X3 [3]
-case-> X4

RHS:

X1 [1]
-nsubjpass-> X3 [3]
-iobj-> X2 [2]
-case-> X5

C: lemma(X4) = FOR; lemma(X5) = ON; [...]

Zapis reguł w XML

```
<rule comment="del amod" equivalent="" id="R197"  
lang="pl en" strength="S" type="Sub">  
  <lhs>  
    <s ref="#197l"/>  
  </lhs>  
  <rhs>  
    <s ref="#197r"/>  
  </rhs>  
  <Constraints/>  
</rule>
```

Zapis reguł w XML

Serializacja ograniczeń w standardzie RuleML:

```
<Constraints>
  <Atom>
    <!--UE({X1})-->
    <Rel>UE</Rel>
    <Set>
      <Var>X1</Var>
    </Set>
  </Atom>
  <Atom>
    <!--intersective(X2)-->
    <Rel>intersective</Rel>
    <Var>X2</Var>
  </Atom>
</Constraints>
```

Zapis reguł w XML

Serializacja drzew zależnościowych w standardzie <tiger2/>:

```
<s xml:id="1971">
  <graph xml:id="1971_g1">
    <terminals>
      <t xml:id="1971_x1" var="X1" ms="" lemma=""
        quantifier="" anchor="1971_a1">
        <edge tiger2:type="dep" label="amod"
          tiger2:target="#1971_x2"/>
      </t>
      <t xml:id="1971_x2" var="X2" ms="" lemma=""
        quantifier="?+" anchor="" />
    </terminals>
  </graph>
</s>
```

Wnioski

Monotoniczność. Wstęp

UE *wszystkie*

Wszystkie hipopotamy umieją tańczyć tango. \Rightarrow

Wszystkie hipopotamy umieją tańczyć.

DE *żaden*

Żaden nosorożec nie lata. \Rightarrow

Żaden nosorożec nie lata szybko.

Intuicje

Operator jest monotonicznie rosnący, gdy pozwala na wnioskowanie z podzbiorów do nadzbiorów. Mówimy, że jest monotonicznie malejący w przeciwnym przypadku.

Monotoniczność. Wstęp

UE *wszystkie*

Wszystkie hipopotamy umieją tańczyć tango. \Rightarrow
Wszystkie hipopotamy umieją tańczyć.

DE *żaden*

Żaden nosorożec nie lata. \Rightarrow
Żaden nosorożec nie lata szybko.

Intuicje

Operator jest monotonicznie rosnący, gdy pozwala na wnioskowanie z podzbiorów do nadzbiorów. Mówimy, że jest monotonicznie malejący w przeciwnym przypadku.

Problemy z oznaczaniem monotoniczności

Przełożenie monotoniczności na własności składniowe i leksykalne jest bardzo złożone i niejednoznaczne.

- ▶ Różne typy fraz wpływają na monotoniczność.
- ▶ Modyfikator nie musi być bezpośrednim rzędnikiem.
Niewiele dzieci aktorów jest grzecznych. ⇒
Niewiele córek aktorów jest grzecznych.
Dzieci niewielu aktorów są grzeczne. ⇒
Córki niewielu aktorów są grzeczne.
- ▶ Operatory modyfikujące monotoniczność mogą być zagnieżdżone.
Np. *Anna nie zapomniała zmusić Marii do wyjścia.*

Wyróżnione obszary

Rozwiązanie: nakładanie ograniczeń na wyróżnione obszary.

- ▶ Reguły czułe na monotoniczność kontekstu przyjmują dodatkowe ograniczenia (UE, DE).
- ▶ Regułę można zastosować tylko wtedy, gdy dane wyrażenie jest w kontekście ustalonym przez regułę.
- ▶ Ograniczenia zachowują minimum informacji potrzebnych z punktu widzenia relacji wynikania

Wyróżnione obszary

I: At most ten **commissioners** spend time at home.

O: At most ten female commissioners spend time at home.

add intersective amod in DE context (en)

LHS:

X1 [1]

RHS:

X1 [1]

-amod-> Y1 [2]

C: intersective(Y1), **DE({X1})**

[...]

Faktywność i quasi-faktywność

Faktywność

Faktywne są czasowniki o takiej treści, że prawdziwość ich dopełnienia zdaniowego wynika z prawdziwości całego zdania złożonego: *Jan wie, że pada. ⇒ Pada.*

Wiele wnioskowań z korpusu RTE opiera się na przekształceniach, które nie są ściśle faktywne, ale tak je traktujemy:

RTE3_d_51.5

I: El-Nashar powiedział, że propaganda przeciwko niemu sprawiła, że ludzie uważają el-Nashara za terrorystę.

O: Propaganda przeciwko niemu sprawiła, że ludzie uważają el-Nashara za terrorystę.

Quasi-faktywność przykłady

RTE3_d_258.2

I: Fiona North, okulistka w SHSSB dodała: obrażenia oczu spowodowane przez sztuczne ognie są wyjątkowo poważne i mogą prowadzić do trwałej ślepoty.

O: Obrażenia oczu spowodowane przez sztuczne ognie są wyjątkowo poważne i mogą prowadzić do trwałej ślepoty.

RTE3_d_609.5

I: Policja oznajmiła, że nie wyklucza się użycia przez policję robota do rozbrajania bomb.

O: Nie wyklucza się użycia przez policję robota do rozbrajania bomb.

Klasyfikacja wnioskowań

- ▶ Sygnatura siły S przyjmuje trzy wartości: S (strong), DF (defeasible), WTF (wishful thinking fallacy)
- ▶ Pozwala ona na powierzchowny podział wnioskowań w zależności od użytych reguł.
- ▶ Tylko reguły S: wnioskowania mocne, przynajmniej jedna DF: wnioskowania odwoływalne, przynajmniej jedna WTF: błąd anotatora.
- ▶ Kroki, w których użyto reguł DF, stanowią zróżnicowany zbiór.

Przykłady DF

RTE3_d_405.10: DF wykonać pierwszy spacer → wylądować

I: Neil Armstrong był pierwszym człowiekiem, który wykonał spacer księżycowy.

O: Neil Armstrong był pierwszym człowiekiem, który wylądował na Księżycu.

RTE3_d_364.8: DF Wirus HIV → AIDS

I: Wirus HIV rozprzestrzenia się w Afryce.

O: AIDS rozprzestrzenia się w Afryce.

RTE3_d_609.7: DF uogólnienie do liczby mnogiej

I: Policja używa robota do rozbrajania bomb.

O: Policja używa robotów do rozbrajania bomb.

Znane ograniczenia

- ▶ Każdy krok w derywacji musi być poprawnym zdaniem danego języka.
- ▶ Jak ludzie rozstrzygają o wynikaniu? Czy przeprowadzają coś w rodzaju derywacji?
- ▶ Reguły i korpusy dziedziczą niektóre z wad oryginalnych zasobów.
- ▶ Ignorujemy wiedzę o świecie: wszystko jest w słowniku.

Podsumowanie

Stworzyliśmy formalizm sankcjonujący relację wynikania na podstawie maksymalnie niskopoziomowych własności tekstu.

Przeprowadziliśmy nowatorski podział par wyników tekstowych ze względu na wiedzę potrzebną do rozpoznania w nich relacji wynikania.

Podział ten umożliwia jakościową ewaluację systemów RTE.

Projekt w liczbach

1025 derywacje (RTE3: 822, FraCaS: 203)

5208 kroki (RTE3: 4639, FraCaS: 569)

167 reguł

14 miesięcy

300 mb danych

19 osób (13 stażystów)

2 konferencje: 20th Amsterdam Colloquium, Szklarska Poręba
Workshop 2016

1 artykuł: A. Przepiórkowski, et al. *Towards a Taxonomy of
Textual Entailments.*

<http://rte.nlp.ipipan.waw.pl>

Podsumowanie

Stworzyliśmy formalizm sankcjonujący relację wynikania na podstawie maksymalnie niskopoziomowych własności tekstu.

Przeprowadziliśmy nowatorski podział par wyników tekstowych ze względu na wiedzę potrzebną do rozpoznania w nich relacji wynikania.

Podział ten umożliwia jakościową ewaluację systemów RTE.

Dziękujemy za uwagę!