Problem ujednoznaczniania sensu w kontekście ekstrakcji relacji semantycznych

Aleksander Pohl

Instytut Podstaw Informatyki Polskiej Akademii Nauk

16 kwietnia 2012

Plan prezentacji

Plan prezentacji

Czym jest ekstrakcja informacji?

Intuicyjnie *ekstrakcja informacji* polega na wydobyciu faktów opisanych (w tekście) w języku naturalny i przekształceniu ich do wybranej reprezentacji (relacyjnej, grafowej, ontologicznej), tak aby mogłoby być przetwarzane automatycznie w systemach komputerowych.

Zadania w obrębie ekstrakcji informacji

Tradycyjnie wyróżnia się następujące zadania w obrębie ekstrakcji informacji (Jurafsky et al., Moens):

- rozpoznawanie wyrażeń nazwowych (*Named Entity Recognition*)
- rozpoznawanie koreferencji (wyrażeń współodnoszących się) (*Coreference Resolution*)
- rozpoznawanie i klasyfikacja relacji semantycznych (*Relation Detection and Classification*)
- rozpoznawanie wyrażeń temporalnych oraz ekstrakcja zdarzeń (*Temporal and Even Processing*)
- wypełnianie szablonów (*Template Filling*)

Przykłady ekstrakcji informacji (1)

Przykłady (D. Jurafsky et al.):

- „*Turing* jest często uznawany za ojca współczesnej informatyki.” – rozpoznanie wyrażenia *Turing* jako odnoszącego się do osoby.
- „*American Airlines* posiadają węzeł lotniczy w *San Juan*.” – rozpoznawanie i klasyfikacja relacji zachodzącej pomiędzy obiektami do których odnoszą się wyrażenia *American Airlines* oraz *San Juan*.

Przykłady ekstrakcji informacji (2)

- „Organizacja Czarny Wrzesień *próbowała zdetonować* trzy samochody pułapki w Nowym Jorku w marcu 1973 roku.” – rozpoznawanie zdarzeń.
- Wypełnianie szablonów:

DETONATION-ATTEMPT	
BOMBER	Czarny Wrzesień
PLACE	Nowy Jork
DATE	marzec 1973
BOMB-COUNT	3
BOMB-TYPE	samochód pułapka

Definicja ekstrakcji informacji – M. F. Moens

*Information extraction is the identification and consequent or concurrent **classification** and **structuring** into semantic classes, of specific information found in **unstructured data sources**, such as natural language text, making the information **more suitable for information processing** tasks.*
– M. F. Moens 2006

Definicja Moens – uwagi

- Sformułowania „nieustrukturyzowane źródła danych” oraz „bardziej zdatne do przetwarzania” są mało precyzyjne.
- Struktura tekstu w języku naturalnym:
  - dokumenty/teksty
  - paragrafy
  - zdania
  - słowa
- Odnalezienie dokumentów zawierających określone **słowa kluczowe** nie stanowi większego problemu z punktu widzenia przetwarzania informacji. Problem pojawia się jeśli np. chcemy odnaleźć zdania posiadające określoną strukturę składniową.

Język przedmiotowy i meta-język – A. Tarski

[...] we have to use two different languages in discussing the problem of the definition of truth and, more generally, any problems in the field of semantics. The first of these languages is the language which is „talked about" and which is the subject matter of the whole discussion; [...]. The second is the language in which we „talk about" the first language [...]. We shall refer to the first language as „the object language", and to the second as „the meta-language".

– A. Tarski 1944, „The Semantic Conception of Truth and the Foundations of Semantics"

Propozycja definicji – A. Pohl

Proces ekstrakcji informacji jest procesem nadawania znaczenia (interpretacji), w którym przechodzi się od opisu danych w terminach meta-języka, do opisu w terminach języka przedmiotowego, dzięki czemu dane źródłowe mogą być bezpośrednio interpretowane w zadaniach przetwarzania informacji.

Problemy związane z ekstrakcją informacji

- „nieprzezroczystość” danych tekstowych
- wszechobecna wieloznaczność: form wyrazowych, struktur składniowych, semantyczna wieloznaczność wyrażeń, itp.
- ograniczona dostępność zasobów językowych/wysoki koszt ich wytworzenia
- wyrażenia wielosegmentowe
- wyrażenia metaforyczne

Plan prezentacji

Wykorzystywane zasoby językowe i źródła wiedzy

- korpusy tekstów:
  - IPI PAN: 250 mln. segmentów
  - notatki PAP AGH: 3,6 mln. segmentów
- słowniki fleksyjne:
  - biblioteka CLP AGH: 138 tys. leksemów
  - Morfologik przekształcony do formalizmu CLP: 270 tys. leksemów
- semantyczne źródła wiedzy:
  - polska Wikipedia, ok. 800 tys. artykułów
  - ontologia Research Cyc, ok. 542 tys. symboli

Korpusy tekstów

- Korpus IPI PAN:
  - wykorzystany jako źródło przykładów uczących przy konstrukcji formalnych szablonów ekstrakcyjnych
  - zalety: język zapytań, tagowanie
  - wady: niewystarczająca wydajność, niezbalansowany
- Korpus PAP AGH:
  - wykorzystany jako źródło przykładów uczących przy konstrukcji semantycznych szablonów ekstrakcyjnych
  - wykorzystany do testowania skuteczności algorytmu
  - zalety: wysoka wydajność[1], wysoka jakość artykułów
  - wady: brak tagowania, brak narzędzi wspomagających, niewielki rozmiar

[1] Po umieszczeniu tekstów w dedykowanej, obiektowej bazie danych ROD.

Słowniki fleksyjne

- CLP AGH:
  - zalety: dobra znajomość rozwiązania, jednoznaczna identyfikacja leksemów posiadających homonimiczne formy bazowe, możliwość zastosowania interfejsu obiektowego, wysoka jakość danych
  - wady: brak wielu popularnych leksemów, brak istotnych relacji morfosyntaktycznych, słaba dokumentacja
- Morfologik:
  - zalety: występowanie wielu leksemów będących składnikami nazw własnych, dość dobrze znany zestaw znaczników bazujących na tagach korpusu IPI PAN
  - wady: średnia jakość danych, brak jednoznacznej identyfikacji leksemów o homonimicznych formach bazowych, konieczność dostosowanie do formalizmu CLP

Wikipedia

- zastosowania:
  - określenie semantycznego powinowactwa wyrażeń
  - ujednoznacznianie wyrażeń
  - rozpoznawanie wyrażeń wielosegmentowych
  - określanie kategorii semantycznej wyrażeń
- zalety: duża ilość reprezentowanych wyrażeń (w szczególności nazw własnych), obecność niejawnej informacji morfologicznej
- wady: niejednorodna jakość materiału, trudności w automatycznym przekształceniu w wysokiej jakości słownik semantyczny

Ontologia ResearchCyc

- zastosowania:
  - źródło par uczących
  - źródło wiedzy na temat relacji generalizacji
  - uogólnianie ograniczeń semantycznych w szablonach ekstrakcyjnych
- zalety: wysokiej jakości dane pozwalające na prowadzenie niezawodnych wnioskowań, duża liczba gotowych do wykorzystania par uczących
- wady: niejednorodne pokrycie obszarów wiedzy, stosunkowo niewielka ilość danych dotyczących wyrażeń będących nazwami własnymi, skomplikowanie, brak (wystarczająco bogatego) polskiego leksykonu

Plan prezentacji

Cel algorytmu ekstrakcji relacji

- *Ponad 10 tys. antylop uciekło z wyjątkowo silnie zaśnieżonych **stepów Mongolii** i przedostało się w poszukiwaniu jedzenia do wschodniej Syberii*
- *stepów* oraz *Mongolii* – rozpoznanie dwóch wyrażeń odnoszących się do obszarów geograficznych
- rozpoznanie relacji *część-całość* zachodzącej pomiędzy wyrażeniami
- chodzi o rozpoznanie *instancji* relacji, tak by określony fragment tekstu mógł zostać oznakowany semantycznie
- nie chodzi o budowanie ontologii
- celem jest tworzenie bazy wiedzy z danymi wyekstrahowanymi z tekstów z zachowaniem informacji o źródle

Koncepcja algorytmu

- cel: budowana szablonów ekstrakcyjnych
- wykorzystanie danych z wielu źródeł wiedzy w celu maksymalizacji liczby rozpoznawanych wyrażeń
- hybrydowy algorytm oparty o przykładu uczące:
  - cechy morfosyntaktyczne określane automatycznie z wykorzystaniem algorytmów uczenia maszynowego
  - cechy semantyczne określane na podstawie ontologii
- ontologia wykorzystywana również do zwiększenia różnorodności przykładów uczących

Struktura algorytmu ekstrakcji relacji

1. wybór relacji (np. całość-część)
2. wygenerowanie par uczących (np. rekin-płetwa)
3. odnalezienie par uczących w korpusie tekstów
4. utworzenie *formalnych* szablonów ekstrakcyjnych (np. *-dat – „płetwa rekina”, „płetwy rekina”, etc.)
5. statystyczna analiza szablonów
6. podział szablonów na grupy (na podstawie odległości argumentów)
7. odnalezienie zdań pasujących do szablonów w korpusie tekstów
8. określenie typu relacji w odnalezionych przykładach
9. uogólnienie ograniczeń semantycznych dla odnalezionych przykładów uczących
10. utworzenie *semantycznych* szablonów ekstrakcyjnych

Kluczowe zadanie dodatkowe

Określenie kategorii semantycznych wyrażeń występujących w tekście:

- rozpoznanie wyrażeń jedno i wielosegmentowych, np. „Spotkanie odbyło się w *Zamku Królewskim*”
- ujednoznacznienie sensu wyrażeń, np.:
  - Zamek Królewski w Warszawie
  - Zamek Królewski na Wawelu
  - Zamek Królewski w Poznaniu
  - . . .
- określenie kategorii semantycznej zdefiniowanej w ontologii Cyc dla rozpoznanych wyrażeń, np. `#$Castle`

Plan prezentacji

Struktura algorytmu ujednoznaczniania sensu

Algorytm ujednoznaczniania wyrażeń oparty o Wikipedię (D. Milne, I. H. Witten 2008)

- rozpoznanie *wyrażeń jednoznacznych*
- określenie *wagi* wyrażeń jednoznacznych na podstawie:
  - powinowactwa semantycznego z pozostałymi wyrażeniami jednoznacznymi
  - statystycznej częstości wykorzystania tych wyrażeń do tworzenia linków do innych artykułów w Wikipedii
- ujednoznacznienie sensu *wyrażeń wieloznacznych* na podstawie drzewa decyzyjnego zbudowanego z wykorzystaniem algorytmu C4.5

Miara powinowactwa semantycznego

Oparta o Wikipedię miara powinowactwa semantycznego (I. H. Witten, D. Milne 2008) wykorzystująca odległość *Google*

$$sr_{google}(a,b) = 1 - \frac{log(max(|A|,|B|)) - log(|A \cap B|)}{log(|W|) - log(min(|A|,|B|))} \quad (1)$$

- $|A|$ – moc zbioru artykułów posiadających odnośniki do artykułu a
- $|A \cap B|$ – moc zbioru artykułów posiadających odnośniki jednocześnie do a i do b
- $|W|$ – moc zbioru wszystkich artykułów w Wikipedii
- w praktyce „miara” ta może być mniejsza od 0!

Określanie wagi artykułów/wyrażeń

Dla każdego jednoznacznego artykułu/wyrażenia:

- w I iteracji: określenie średniego powinowactwa semantycznego z pozostałymi artykułami
- w II iteracji: określenie wagi artykułu jako średniej arytmetycznej powinowactwa semantycznego z pozostałymi artykułami oraz miary *link probability*

link probability – częstość z jaką określone wyrażenie, które reprezentuje artykuł jest wykorzystywane w Wikipedii jako odnośnik do innych artykułów

Cechy wykorzystane do indukcji drzewa decyzyjnego

- średnia ważona *powinowactwa semantycznego* artykułu (reprezentującego sens wyrażenia) z pozostałymi artykułami (*relatedness*)
- prawdopodobieństwo określonego sensu, obliczone jako proporcja odnośników, których treść stanowi dane wyrażenie, prowadzących do danego artykułu w stosunku do liczby wszystkich odnośników zarejestrowanych dla tego wyrażenia (*sense probability*)
- „jakość” kontekstu danego wyrażenia, określona jako suma wag wyrażeni określonych wcześniej (*goodness*)

Przykłady uczące

Indykcja drzewa decyzyjnego odbywa się na podstawie artykułów Wikipedii:

- wybierane są artykuły zawierające odpowiednią liczbę odnośników wewnętrznych (w oryginalnym artykule > 50)
- z artykułów ekstrahowane są pary
  - treść odnośnika – wyrażenie, np. „*jądro systemu operacyjnego* charakteryzowało się...”
  - cel odnośnika – artykuł Wikipedii, np. *Jądro systemu*
- dla tej pary obliczane są cechy przedstawione wcześniej, stanowi ona pozytywny przykład uczący
- negatywne przykłady uczące generowane są na podstawie wszystkich pozostałych artykułów, do których tworzone są odnośniki o tej samej treści

Wyniki działania algorytmu

Tablica: Skuteczność algorytmu Milne i Wittena

	precision	recall	f-measure
Losowy sens	50,2	56,4	53,1
Najczęstszy sens	89,3	92,2	90,7
Milne i Witten	98,4	95,7	97,1

Uwagi:

- w eksperymencie użyto około 1 mln. przykładów uczących
- wyniki uwzględniają również wyrażenia jednoznaczne (!)
- wyniki są obliczane na podstawie ewaluacji wewnętrznej (tzn. na podstawie danych wygenerowanych z Wikipedii)
- założona jest dostępność dużej liczby ujednoznacznionych wyrażeń (ponad 50)

Wprowadzone modyfikacje

- użycie innej miary powinowactwa semantycznego – Jaccard
- wprowadzenie dodatkowych cech w procesie uczenia
- ewaluacja uwzględniająca wyłącznie wieloznaczne odnośniki
- bardziej realistyczne dane testowe (5-100 odnośników)
- ewaluacja zrealizowana również dla polskiej Wikipedii

Ulepszona miara

Podobnie jak w Wikipedia Minerze miara wykorzystuje informację dotyczącą odnośników do artykułów, oparta jest jednak na szeroko wykorzystywanej mierze Jaccarda:

$$sr_{jaccard}(a,b) = \begin{cases} \frac{1}{1-log\left(\frac{|A\cap B|}{|A\cup B|}\right)} & |A\cap B| > 0 \\ 0 & a \neq b \\ 1 & a = b \end{cases} \quad (2)$$

- $|A|$ – moc zbioru artykułów posiadających odnośniki do artykułu a
- $|A \cap B|$ – moc zbioru artykułów posiadających odnośniki jednocześnie do a i do b
- $|A \cup B|$ – moc zbioru artykułów posiadających odnośniki do a lub do b

Dodatkowe cechy

- *pozycja* artykułu obliczona na podstawie ważonego *powinowactwa semantycznego* (*relatedness position*)
- *pozycja* sensu obliczona na podstawie prawdopodobieństwa jego wystąpienia (*sense position*)
- *link probability* – określone wcześniej

Wyniki dla języka angielskiego

Tablica: Porównanie skuteczności dla języka angielskiego

	precision	recall	F_1-measure
Losowy sens	39.1	20.8	27.2
Losowy sens o P > 0.5%	44.2	45.1	44.6
Najczęstszy sens	82.8	84.6	83.7
sr_G	83.5	84.4	84.0
sr_G + nowe cechy	83.3	85.0	84.1
sr_J	87.2	93.0	90.0
sr_J + nowe cechy	**90.5**	**94.4**	**92.4**

Uwagi:

- liczba przykładów uczących: 3 mln.
- liczba przykładów testowych: 1 mln

Wyniki dla języka polskiego

Tablica: Porównanie skuteczności dla języka polskiego

	precision	recall	F_1-measure
Losowy sens	39.7	26.4	31.7
Losowy sens o P > 0.5%	47.0	47.3	47.2
Najczęstszy sens	81.6	82.2	81.9
sr_G	82.5	83.5	83.0
sr_G + nowe cechy	84.9	83.2	84.0
sr_J	85.4	89.8	87.6
sr_J + nowe cechy	**90.4**	**93.0**	**91.7**

Uwagi:

- liczba przykładów uczących: 1,16 mln.
- liczba przykładów testowych: 390 tys.

Dziękuję!