Rozpoznawanie relacji sematycznych w tekscie, porzez klasyfikację kontekstu relacji przy pomocy klasyfikatora **CRF**

Jakub A. Gramsz
Michał Krautforst

24 stycznia 2014

Spis treści

1 Wstęp **2**
- 1.1 Klasyfikator CRF . . . 2
  - 1.1.1 Motywacja . . . 2
  - 1.1.2 Charakterystyka klasyfikatora . . . 2
- 1.2 Rozpoznanie literatury . . . 2
  - 1.2.1 “Conditional random fields: Probabilistic models for segmenting and labeling sequence data” [2] . . . 2
  - 1.2.2 “Conditional random fields: An introduction” [4] . . . 3
  - 1.2.3 “Document Summarization Using Conditional Random Fields” [3] . . . 3
  - 1.2.4 “Integrating probabilistic extraction models and data mining to discover relations and patterns in text” [1] . . . 3

2 Praca badawcza **4**
- 2.1 Implementacja . . . 4
- 2.2 Badania . . . 4
  - 2.2.1 Zbiór uczący . . . 4

3 Wyniki i wnioski **5**
- 3.1 Wyniki . . . 5
- 3.2 Wnioski . . . 5

1

Wstęp

1.1 Klasyfikator CRF

1.1.1 Motywacja

Klasyfikator CRF utworzony został pod kątem segmentacji i znakowania danych sekwencyjnych [2]. Klasyfikator ten często wykorzystywany w przetwarzaniu języka naturalnego [1, 3] , jest stosunkowo młodą metodą klasyfikacji, wykazująca potencjał do badań. Klasyfikator ten osiągną wysoką jakością klasyfikacji przy oznaczaniu nazw własnych w tekstach.

1.1.2 Charakterystyka klasyfikatora

Klasyfikator CRF został zaproponowany, aby przetwarzać sekwencje w obu kierunkach [2], co było niemożliwe stosując ukryte modele Markowa. Klasyfikator CRF poszukuje najbardziej prawdopodobnej sekwencji przyporządkowań etykiet do sekwencji danych, przy czym składowe prawdopodobieństwa dla poszczególnych elementów w sekwencji wyznaczane są zarówno na podstawie bezpośrednio poprzedzającego elementu oraz bezpośrednio następującego, co odróżnia tę metodę od ukrytych modelów Markowa (HMM). W ogólności metoda ta zakłada, że może pracować nie tyko na łańcuchu (sekwencji), ale na dowolnym grafie skierowanym (wtedy mamy więcej elementów bezpośrednio poprzedzających i następujących), jednakże zazwyczaj w literaturze zakłada się pracę na łańcuchu, czyli sekwencji (nawet w pracy prezentującej CRF [2]).

1.2 Rozpoznanie literatury

1.2.1 “Conditional random fields: Probabilistic models for segmenting and labeling sequence data” [2]

Artykuł w którym zaproponowano nowy klasyfikator (ang. Conditional random fields) służący do segmentacji lub etykietowania danych sekwencyjnych mający w tych zadaniach pokonywać wady ukrytych łańcuchów Markowa i gramatyk stochastycznych. Prezentowany klasyfikator przedstawiany jest również jako rozwiązanie ograniczeń modelów Markowa maksymalizujących entropię (MEMM) oraz innych modeli tego typu bazujących na grafach skierowanych.

W artykule dokonano porównania efektywności oraz jakości etykietowania danych sekwencyjnych za pomocą CRF oraz HMM i MEMM. Wskazuje on na dużo lepsze wyniki uzyskane przez prezentowaną metodę CRF.

1.2.2 "Conditional random fields: An introduction" [4]

Artykuł opisujuje sposób działania klasyfikatora CRF.

1.2.3 "Document Summarization Using Conditional Random Fields" [3]

Artykuł opisujący wykorzystanie CRF do zadania streszczania dokumentów. Opisuje zastosowanie klasyfikatora do problemu z dziedziny inżynierii języka naturalnego, wskazując mi. cechy wykorzystane do opisania danych.

1.2.4 "Integrating probabilistic extraction models and data mining to discover relations and patterns in text" [1]

Przedstawienie zadania odnajdywania relacji w tekście jako problem etykietowania danych sekwencyjnych z wykorzystaniem CRF.

2

Praca badawcza

2.1 Implementacja

Poniżej przedstawiamy zadania.

1. Wydobycie par hiponim-hiperonim o zadanej odległości między sobą wprost z bazy Słowosieci.
2. Budowa grafu relacji hiperonimi.
3. Oznaczenie otrzymanych par w zadanym korpusie.
4. Podział korpusu na zbiór uczący oraz testujący.
5. Wydobycie oznaczonych kontekstów.
6. Wypisywanie przykładów false positive
7. Dobór cech dla klasyfikatora **CRF**.
8. Wyznaczenie odległości pomiędzy parami oznaczonymi poprzez klasyfikator (pośrednio oznaczonymi poprzez kontekst).

2.2 Badania

2.2.1 Zbiór uczący

Zbiór uczący stanowią zdania pochodzące z korpusów języka polskiego z oznaczonymi kontekstami par słów znajdujących się w relacji hiperonimi znajdujących się w różnych odległościach od siebie (1 do 3) w słowosieci. Każdy token opisany wektorem cech.

- korpus wykorzystany w testach - Korpus Języka Polskiego Politechniki Wrocławskiej
- długości kontekstów 5, 10, 15
- pary hiponim-hiperonim o odległościach 1, 2, 3.

3

Wyniki i wnioski

3.1 Wyniki

kontekst	Odl. par	Precyzja	Kompletność	F-score
5	1	72.73%	18.18%	29.09%
15	1	69.57%	17.02%	27.35%
5	2	78.57%	14.86%	25.00%
10	2	70.83%	15.18%	25.00%
10	3	63.89%	15.44%	24.86%
5	3	80.00%	13.79%	23.53%
10	1	73.33%	13.25%	22.45%
15	2	48.65%	13.24%	20.81%
15	3	34.38%	7.10%	11.76%

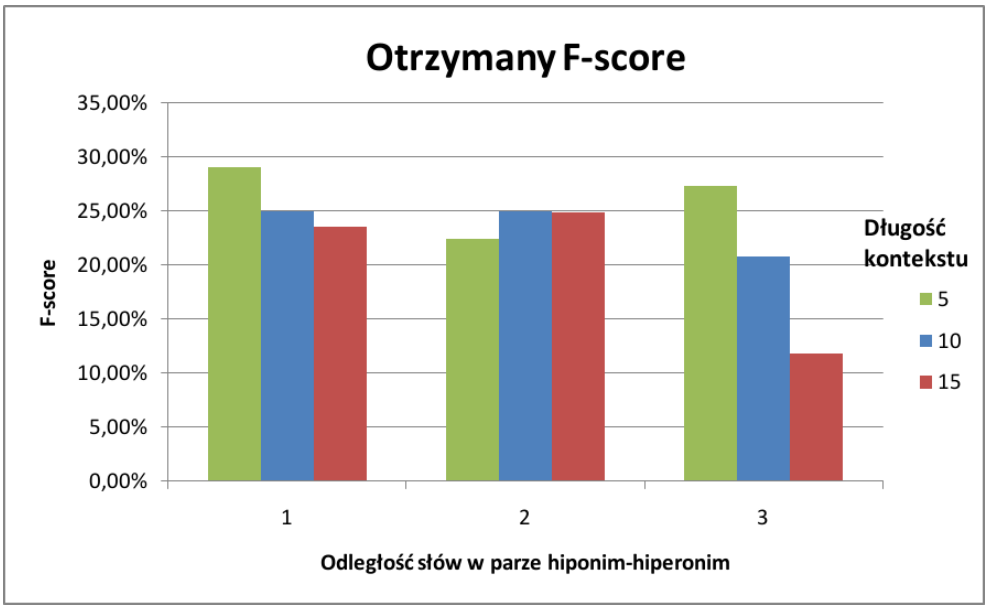
Tablica 3.1: Podsumowanie najlepszych wyników z badań jakości klasyfikatora względem długości kontekstu oraz odległości między leksemami w sensie relacji hiperonimi.

3.2 Wnioski

- Badania prowadzone były na stosunkowo małym korpusie, warto przeprowadzić analogiczne testy dla większej ilości danych, napotkano jednak problem w postaci bardzo dużego zapotrzebowania na zasoby pamięci.
- Skorzystanie z par hiponim-hiperonim o większej odległości powoduje spadek precyzji bez zmian kompletności.
- Długość kontekstu wpływa na liczbę wygenerowanych przykładów pozytywnych, jednak wraz z jego wzrostem spada jakość klasyfikacji.
- Manipulacja cechami niezauważalnie wpływa na wyniki (cech powinno być możliwie dużo).
- Zmniejszenie liczby negatywnych przypadków pogarsza jakość klasyfikacji. Należałoby zrównoważyć proporcje liczby przypadków pozytywnych i negatywnych, jednak trudne jest określenie kryterium.

Dł. Kontekstu	Odl. Par	Walidacja	TP	FP	FN	Precyzja	Kompletność	F-score
5	1	7	25	3	110	89.29%	18.52%	30.67%
5	1	10	8	3	36	72.73%	18.18%	29.09%
15	1	10	16	7	78	69.57%	17.02%	27.35%
10	2	10	17	7	95	70.83%	15.18%	25.00%
5	2	10	11	3	63	78.57%	14.86%	25.00%
10	3	10	23	13	126	63.89%	15.44%	24.86%
5	1	5	33	0	202	100.00%	14.04%	24.63%
5	2	5	51	10	313	83.61%	14.01%	24.00%
5	2	7	31	15	186	67.39%	14.29%	23.57%
5	3	10	12	3	75	80.00%	13.79%	23.53%
10	3	7	61	34	370	64.21%	14.15%	23.19%
5	3	5	60	17	389	77.92%	13.36%	22.81%
5	3	7	37	13	239	74.00%	13.41%	22.70%
10	1	10	11	4	72	73.33%	13.25%	22.45%
10	2	7	48	28	309	63.16%	13.45%	22.17%
15	2	10	18	19	118	48.65%	13.24%	20.81%
15	1	7	35	20	262	63.64%	11.78%	19.89%
10	1	7	24	12	224	66.67%	9.68%	16.90%
10	3	5	70	73	671	48.95%	9.45%	15.84%
10	1	5	34	13	366	72.34%	8.50%	15.21%
15	1	5	42	25	446	62.69%	8.61%	15.14%
15	3	7	43	69	425	38.39%	9.19%	14.83%
15	2	7	32	38	358	45.71%	8.21%	13.91%
15	3	5	58	103	726	36.02%	7.40%	12.28%
15	3	10	11	21	144	34.38%	7.10%	11.76%
15	2	5	36	58	639	38.30%	5.33%	9.36%

Tablica 3.2: Wszystkie wyniki otrzymane w wyniku badań jakości klasyfikatora względem długości kontekstu oraz odległości między leksemami w sensie relacji hiperonimi.



Wykres 3.1: Porównanie F-score klasyfikatorów w zależności od długości kontekstów oraz odległości pomiędzy hiponimem a hiperonimem w parach.

- Klasyfikacja cechuje się wysoką precyzją, mimo zróżnicowanych przypadków pozytywnych oraz rzadkiej powtarzalności kontekstów.
- Zaproponowana metoda odnajdowania relacji semantycznych w tekście poprzez oznaczanie ich kontekstów, nie przynosi zadowalających rezultatów, niemniej jednak może stanowić dobry wstęp na podstawie, którego innymi metodami można by proces ten kontynuować. Przykładowo oznaczone konteksty można by grupować na podstawie ich podobieństwa a następnie wyciągać z nich bardziej ogólne wzorce.

—

Wykaz zadań wraz z przynależnością do członków grupy:

Michał Krautforst	— skrypt to wyciągania par hiponim-hiperonim dla zadanej odległości — skrypt oznaczający wyciągnięte pary w zadanym korpusie — skrypt dzielący oznaczony korpus na podzbiór uczący i testujący — skrypt co wyciągania oznaczonych kontekstów — przeprowadzenie badań końcowych
Jakub Gramsz	— skrypt to wyciągania par hiponim-hiperonim dla zadanej odległości — skrypt do sprawdzania odległości wyodrębnionych par — badanie wpływu doboru cech — graf relacji hiperonimi (jednak generowanie par — na jego podstawie działa za wolno)

Bibliografia

[1] A. Culotta, A. McCallum, and J. Betz. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 296–303. Association for Computational Linguistics, 2006.

[2] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[3] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen. Document summarization using conditional random fields. In *IJCAI*, volume 7, pages 2862–2867, 2007.

[4] H. M. Wallach. Conditional random fields: An introduction. *Technical Reports (CIS)*, page 22, 2004.