IDENTYFIKACJA PRZEDMIOTÓW OPINII W JĘZYKU POLSKIM

Aleksander Wawer

axw@ipipan.waw.pl

Agenda

- **Wstęp**
- **Anotacja i zbiory danych**
  - **Treebanki: recenzje, podzbiór składnicy**
  - **Twitter**
- **Identyfikacja przedmiotów opinii**
  - **Wzorce zależnościowe**
  - **Wzorce zależnościowe + CRF**
  - **Gramatyka powierzchniowa**
- **Konkluzje**

Przedmioty opinii (opinion targets)

- Przedmioty opinii to słowa lub wyrażenia, względem których skierowany lub wyrażony jest wydźwięk, opinia.
  - W recenzjach produktów są to atrybuty, aspekty lub obiekty.
    Bateria tego laptopa jest fatalna, ale ekran ma świetne kolory.
    Ten Dell jest świetny.
- Przedmioty opinii mogą denotować wiele typów obiektów – wszystko, względem czego chcemy wyrazić opinię!
- Mogą obejmować nie tylko obiekty, ale też wydarzenia, stany, przekonania, decyzje lub czyny innych osób.
  - Nawet opinie mogą stać się przedmiotami opinii:
    Nie podoba mi się Twoja negatywna opinia na ten temat.

Zbiór danych: recenzje

- Zdania pochodzące ze zbioru recenzji, pobranych w ~2011 (projekt Nekst) z jednego z największych portali agregujących opinie
  - Dwa typy produktów: perfumy i ubrania

	perfume	clothes
sentences	946	418

- Manualne oznaczenia
  - Poprawności obydwu słów:
    - czy rzeczywiście wyrażają wydźwięk / nie są neutralne,
    - czy rzeczywiście są przedmiotami opinii.
  - Czy zachodzi związek pomiędzy nimi (czy wyrażenie wydźwięku wyrażone jest względem tego właśnie przedmiotu opinii)?

Zbiór danych: twitter

- 1000 tweetów, wylosowanych z danych zbieranych w projekcie Trendminer project (www.trendminer.eu)
  - Tweety pisane zazwyczaj przez dziennikarzy i polityków
- Przefiltrowane pod kątem wystąpienia przynajmniej jednego znanego słowa z wydźwiękiem (http://zil.ipipan.waw.pl/SlownikWydzwieku)
- Dla tego zbioru (300 tweetów) powstała anotacja wydźwięku i przedmiotów opinii
- Wśród nich, jedynie 122 zawiera co najmniej jedną krotkę wydźwięk+przedmiot opinii

Zbiór danych: podzbiór Składnicy

- Do eksperymentów z przedmiotami opinii wykorzystana została wersja zależnościowa Składnicy
- Przefiltrowane pod kątem wystąpienia przynajmniej jednego znanego słowa z wydźwiękiem (http://zil.ipipan.waw.pl/SlownikWydzwieku)
- W ten sposób wybranych zostało ok. 2000 zdań
- Losowo wybrana połowa tego zbioru (1000 zdań) trafiła do anotacji wydźwiękiem i przedmiotami opinii

Ekstrakcja przedmiotów opinii oparta o wzorce składniowe

- Hipoteza: słownik wydźwięku (lub dowolna metoda znakowania wydźwięku na poziomie wyrazowym lub frazowym) + informacja o przechodzeniu drzewa zależnościowego (zbiór wzorców) są wystarczające dla rozpoznawania przedmiotów opinii.
- Zaczynamy od słowa z wydźwiękiem (S), następnie przez sekwencję ruchów, opisanych zgodnie z pewnym systemem formalnym, dochodzimy do przedmiotu opinii – (słowa T).
- Jak opisać tego typu wzorce?
  - Niemiecki TIGERSearch
  - Stanford SemGrex, modyfikacja Tregex

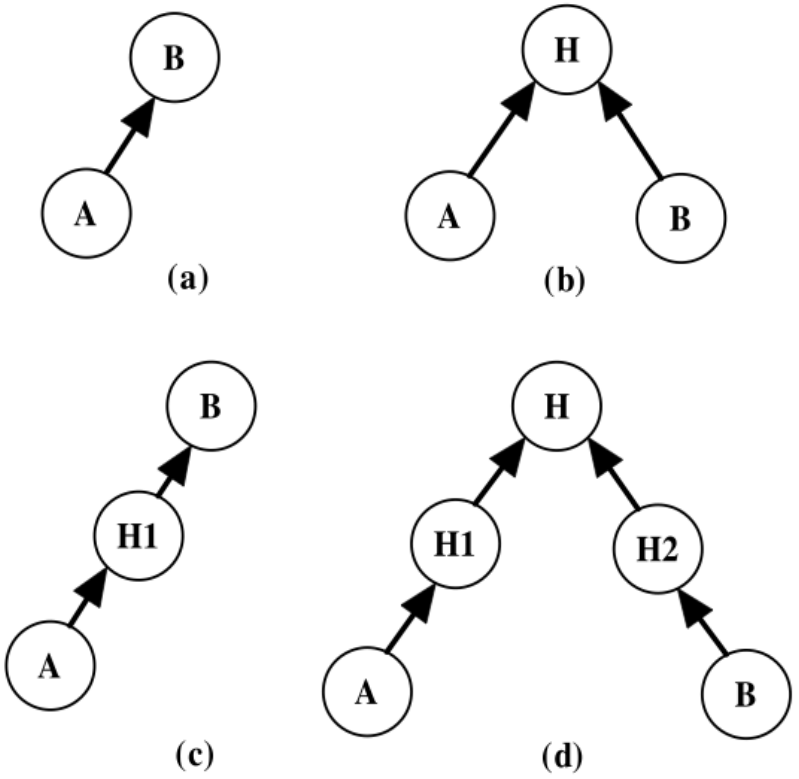
Prehistoria

- “Opinion Word Expansion and Target Extraction through Double Propagation”, Qiu et al. 2011

Wydźwięk: przymiotniki (A)

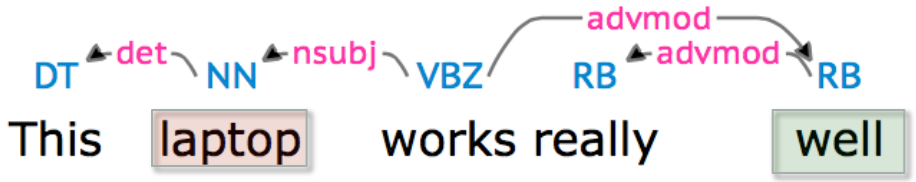
Przedmioty: rzeczowniki (B)

Jak wyrazy z wydźwiękiem (A) łączą się z przedmiotami opinii (B)



Wzorce zależnościowe

- Opisują strukturę zdaniową pomiędzy słowami z wydźwiękiem (opiniami) a przedmiotami opinii
- Tokeny wyrażone jako [..]
- Relacje zależnościowe jako < lub >, zależnie od kierunku
- Możemy określić części mowy lub lematy, które mają dopasować się do tokenów (np. [pos:verb] pasuje do czasowników).
- Możemy określić typ relacji zależnościowej.



[pos:NN] nsubj< [pos:VBZ] >advmod [pos:RB]

Ekstrakcja przedmiotów opinii oparta o wzorce składniowe

- Dla każdej pary S-T w korpusie, generujemy opis ścieżki, zaczynając od S, aż do słowa T.
- Wzorce czytamy od lewej (S) do prawej (T)
- Pierwsza wersja: części mowy oraz etykiety krawędzi zależnościowych
- W ten sposób, korzystając z korpusu recenzji, wygenerowanych zostało 173 wzorców
  - 20-30 relatywnie powtarzalnych wzorców oraz “długi ogon” wzorców opisujących pojedyncze zdania.
  - Precyzja tego zbioru na banku recenzji to 0.73.

Wzorce zależnościowe: ekstrakcja przedmiotów opinii

Najczęstsze 11 wzorców (dopasowanie do co najmniej 20 zdań)

path	precision	matched	not matched	total
[pos:adj] <adjunct [pos:subst]	0.886	396	51	447
[pos:fin] >comp [pos:prep] >comp [pos:subst]	0.814	48	11	59
[pos:adj] >adjunct [pos:prep] >comp [pos:subst]	0.906	48	5	53
[pos:adj] <adjunct [pos:subst] >adjunct [pos:prep] >comp [pos:subst]	0.333	16	32	48
[pos:adj] <pd [pos:fin] >subj [pos:subst]	0.909	40	4	44
[pos:adj] <adjunct [pos:subst] <conjunct [pos:conj] >conjunct [pos:subst]	0.333	11	22	33
[pos:adj] <conjunct [pos:interp] <adjunct [pos:subst]	0.939	31	2	33
[pos:fin] >adjunct [pos:prep] >comp [pos:subst]	0.433	13	17	30
[pos:adj] <adjunct [pos:subst] >adjunct [pos:subst]	0.64	16	9	25
[pos:adj] <conjunct [pos:conj] >conjunct [pos:subst]	0.625	15	9	24
[pos:fin] <conjunct [pos:conj] >conjunct [pos:fin] >subj [pos:subst]	0.304	7	16	23

Wzorce zależnościowe: ograniczenia

- Wzorce są użyteczną metodą ekstrakcji przedmiotów opinii, istnieje więcej wzorców niż tylko te opisane w [Qiu et al, 2011]
- Czy wzorce są wystarczająco dobrą metodą?
- Możliwości poprawienia obejmują m.in.
  - Odkrywanie przedmiotów opinii **nie wskazywanych** przez żaden ze znanych wzorców, tylko na podstawie wnioskowania statystycznego opartego o kontekst
  - Odkrywanie przedmiotów opinii **wskazywanych** przez znany wzorzec, podniesienie precyzji.
- Propozycja rozwiązania: wprowadzenie drugiego kroku
  - tagger CRF, korzystający z m.in. ze wzorców.

Wzorce zależnościowe + CRF: opis cech

- W (..) zapisujemy pozycję, 0 to bieżący token
- Przykładowo **(-1,0,1**) jest oknem złożonym z poprzedniego, bieżącego i następnego tokenu
- [lemma]: unigramy lematów na (-1,0,1), bigramy w (-1,0), (0,1);
- [POS]: unigramy, bigramy i trigramy części mowy w (-2,-1,0,1,2);
- [dep]: unigramy i bigramy etykiet relacji zależnościowych w (-2,-1,0,1,2);
- [ruleAny]: czy bieżący token jest wskazywany jako T-słowo przez jakikolwiek znany wzorzec zależnościowy
- [ruleID]: identyfikator wzorca zależnościowego wskazującego na bieżący token
- [S]: czy bieżące słowo posiada wydźwięk inny niż neutralny

Wzorce zależnościowe + CRF: wyniki

template	description	tPrec	tRec	tF1	mPrec	mRec	mF1	features
T1	[lemma]	0.586	0.33	0.421	0.768	0.656	0.693	17435
T2	[lemma]+[POS]+[dep]	0.553	0.466	0.505	0.756	0.719	0.735	25234
T3	[POS]+[dep]	0.548	0.426	0.478	0.752	0.699	0.721	7805
T4	[POS]+[dep]+[ruleAny]	0.783	0.891	0.833	0.887	0.936	0.91	7808
T5	[POS]+[dep]+[ruleAny]+[ruleID]	0.823	0.901	0.859	0.908	0.943	0.924	8048
T6	[POS]+[dep]+[ruleAny]+[ruleID]+[S]	0.829	0.889	0.857	0.91	0.937	0.923	8067

Wyniki rozpoznawania przedmiotów opinii w 10-krotnej walidacji krzyżowej na zbiorze recenzji

Miary:

- Mikro precyzja (tPrec), recall (tRec) i F1 (tF1),
- Makro precyzja (mPrec), recall (mRec) i F1 (mF1).

Wzorce zależnościowe + CRF: wyniki

- Leksykalna przestrzeń cech (T1) nie działa dobrze
  - Lematy nie są efektywną metodą ekstrakcji przedmiotów opinii
- Czysto składniowa przestrzeń cech (T3) jest również zła
- Wprowadzenie wzorców zależnościowych ma znaczący pozytywny wpływ (T4-T6), precyzja powyżej 0.7
- Optymalny zestaw cech zawiera identyfikator wzorca zależnościowego, wskazującego na bieżący token
- Najlepsze przestrzenie cech nie potrzebują więcej niż 8 tys. cech

Identyfikacja przedmiotów opinii metodą powierzchniową

- Reguły gramatyki są ogólne (np. fraza rzeczownikowa), nie stworzone z myślą o zadaniu łączenia wydźwięku i przedmiotów opinii.
  - Używam gramatyki Spejd tworzonej w ramach NKJP (Głowińska et al.)

- Założenie: wyrażenie wydźwięku jest związane z przedmiotem opinii, jeśli oba występują w ramach tej samej grupy syntaktycznej.
  - Czy każdej? Czy istnieje związek z typem grupy?

Gramatyka powierzchniowa: recenzje

Rule Name	TP	FP
Adj + Noun	173	0
NGg: Noun + n-Noun in gen	53	24
NGa: 2*Adj + Noun	17	5
NGa: Adj + Noun	0	61
NGa: Adj + Noun + Adj	10	3
NGa: Noun + Adj	45	18
NGk: N|NG i N|NG (coordination)	22	8
AdjGe: one of ...	9	7
PrepNG: Prep + NG	101	35
PrepAdjG	2	21
AdvG: Adv + Adv	0	15
AdjGk: Adj i Adj (coordination)	0	19
AdjG: Adv + Adj	0	81
CG: subordinate clause with że, żeby (1)	26	12
...	...	...
Total	467	325

- TP – true positive
- FP – false positive

Gramatyka powierzchniowa: tweety

Rule Name	TP	FP
NGx: pronoun + Adj gen	1	1
NGg: Noun + n-Noun w gen	7	2
NGg: Noun + n-Noun (gen)	1	0
NGs: Noun + n-Noun (nom)	0	1
NGa: Adj + Noun	21	6
NGk: N|NG i N|NG (coordination)	1	0
PrepNG: Prep + NG	9	6
PrepNG with a group in quotes	1	1
NG with adjunct in nom (1)	1	0
PrepAdjG	0	1
AdjG: 2*Adj	0	1
AdjG: Adv + Adj	0	5
AdvG: Adv + Adv	0	3
CG: subordinate clause with że, żeby (2)	1	0
CG: subordinate clause with że, żeby (1)	2	0
Total	45	27

- TP – true positive
- FP – false positive

Wzorce zależnościowe: tweety

Rule Name	TP	FP
[pos:adj] <pd [pos:fin] >subj [pos:subst]	1	4
[pos:adj] <adjunct [pos:subst]	16	8
[pos:adj] <adjunct [pos:subst] >app [pos:subst]	1	8
[pos:adj] <adjunct [pos:subst] >adjunct [pos:subst]	0	2
[pos:adj] <adjunct [pos:subst] <adjunct [pos:subst]	0	1
[pos:adv] <comp [pos:fin] >subj [pos:subst]	1	1
[pos:adv] <adjunct [pos:fin] >subj [pos:subst]	0	1
[pos:ppas] <adjunct [pos:subst]	2	1
[pos:fin] >obj [pos:subst] >adjunct [pos:subst]	0	1
[pos:fin] >subj [pos:subst] >app [pos:subst]	0	2
[pos:fin] <adjunct [pos:subst]	0	1
[pos:subst] <adjunct [pos:subst]	0	1
[pos:subst] <comp [pos:prep] <adjunct [pos:subst]	0	2
[pos:subst] <obj [pos:fin] >subj [pos:subst]	0	3
[pos:adj] <conjunct [pos:conj] >conjunct [pos:fin] >subj [pos:subst]	0	2
[pos:adj] <adjunct [pos:subst] <obj [pos:fin] >subj [pos:subst]	0	1
[pos:adj] <adjunct [pos:subst] <comp [pos:prep] <adjunct [pos:subst]	0	4
[pos:adj] <adjunct [pos:subst] <conjunct [pos:conj] >conjunct [pos:subst]	0	1
Total	21	44

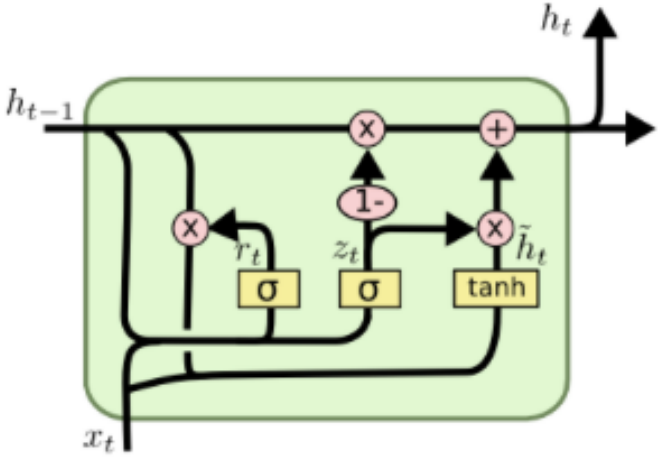
Porównanie metod

	Powierzchniowa		Zależnościowa (*)	
	precyzja	recall	precyzja	recall
twitter	0.62	0.47	0.32	0.27
recenzje	0.59	0.47	0.73	n/a (*)
			0.83 (+CRF)	0.89 (+CRF)

* wzorce były wygenerowane na tym samym zbiorze

Krótki eksperyment z głębokim uczeniem

Sieć GRU (Gated Recurrent Unit), wariant RNN i LSTM

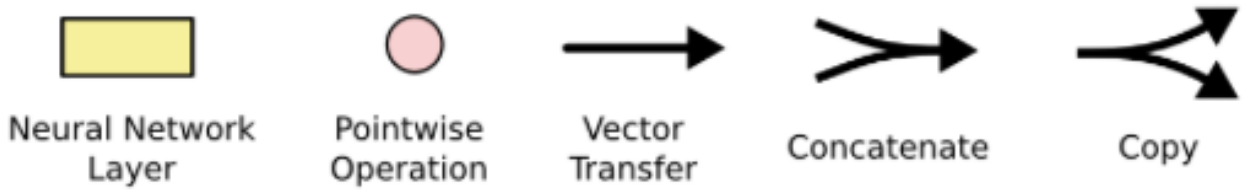


$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$



Porównanie metod

	Powierzchniowa		Zależnościowa (*)	
	precyzja	recall	precyzja	recall
twitter	0.62	0.47	0.32	0.27
recenzje	0.59	0.47	0.73	n/a (*)
			0.83 (+CRF)	0.89 (+CRF)

* wzorce były wygenerowane na tym samym zbiorze

Sieci GRU			
Trenowane	Testowane	precyzja	recall
recenzje	składnica	~0.3	~0.1
recenzje /90%	recenzje /10%	~0.6	~0.3

Konkluzje

- Eksperymenty obejmowały
  - Dwie metody przetwarzania składniowego, powierzchniową i zależnościową Zbiory danych: 122 tweety i >1000 zdań z recenzji (+podzbiór Składnicy)
- Wnioski z obserwacji danych
  - Grupy nominalne z rzeczownikiem i przymiotnikiem są najczęstszym sposobem wiązania wydźwięku i przedmiotów opinii
  - Istnieje jednak kilka innych, istotnych i całkiem licznych wzorców składniowych
- Gramatyka powierzchniowa ma podobną skuteczność na obu typach danych, tweetach i recenzjach
- Na tweetach jakość metody zależnościowej jest zauważalnie gorsza
  - Przypuszczalny powód: niska jakość parsowania
  - Precyzję można podnieść
    - Wybrać tylko niektóre, pasujące grupy syntaktyczne i zależnościowe
    - Dodać uczenie maszynowe jako kolejny krok (vide CRF), “możliwy bag of words”
- Metoda zależnościowa (wzorce) jest skuteczniejsza na recenzjach

Konkluzje

- Eksperymenty z dodaniem CRF jako kolejnego kroku przetwarzania po wzorcach zależnościowych
  - Wzorce użyte jako część przestrzeni cech CRFa
  - **Podniesienie precyzji** względem metody samych wzorców (niektóre z nich zwracają sporo FP)
  - Nieco zaskakujący efekt: najlepiej funkcjonujące przestrzenie cech nie są oparte o leksykę!
- (Względna) **niezależność względem dziedziny** i warstwy leksykalnej!

Wnioski na przyszłość

- Nadal nierozwiązany problem to fakt częściowego tylko wpływu składni
  - Problem “długiego ogona”: często występujące wzorce składniowe o względnie wysokiej precyzji obejmują tylko 60% krotek S-T
- Przypuszczalnie najlepszy schemat postępowania
  1. Reguły składniowe identyfikują potencjalne przedmioty opinii (wysoki recall)
  2. Uczenie maszynowe (podnosi precyzję)
- Głębokie uczenie maszynowe – obiecujące?
  - Problemem jest brak kontroli nad tym, jakie elementy wektorów (embeddings) są wykorzystane w trenowaniu, nadmierna leksykalizacja

Dziękuję za uwagę!

- Treebank recenzji i treebank-podzbiór Składnicy (tweety?)
- Wzorce zależnościowe
- Narzędzie OPFI

można pobrać z

http://zil.ipipan.waw.pl/OPTA