Porównywanie tagerów dopuszczających niejednoznaczności

(na przykładzie tagerów wykorzystanych w Korpusie IPI PAN)

Danuta Karwańska

3 listopad 2008

Plan prezentacji

1. Wprowadzenie
   - Problem niejednoznaczności
   - Poprawna interpretacja
2. Metody ewaluacji
   - Omówienie istniejących miar
   - Nowy sposób analizowania wyników
3. Skuteczność tagerów wykorzystywanych w korpusie IPI PAN
   - Omówienie

Intepretacje morfosyntktyczne

Każdemu **segmentowi** w tekście możemy przypisać **znacznik** (lub, w wypadku segmentów o kilku możliwych interpretacjach, kilka znaczników) **morfosyntaktyczny** interpretujący dany segment np.:

a *„Pamiętam go **pijanego**” (Biernik)*

b *„Widziałam ją **pijaną**.” (Biernik, Narzędnik)*

Oznaczenia

Wprowadźmy następujące oznaczenia:

- s - segment, t - intepretacja morfosyntaktyczna,
- S_s - zbiór wszystkich segmentów w danym tekście
- $S_{(s,t)}$ -zbiór wszystkich par segment - interpretacja,
- $I_X(s)$ - zbiór intepretacji morfosyntaktycznych przypisanych segmentowi s przez metodę X,
- $i_X(s,t) = \begin{cases} 1 & \textit{jeśli } t \in I_X(s) \\ 0 & \text{wpp} \end{cases}$,
- ZS - złoty standard
- Ti - tager Ti

Przypisywanie segmentom intepretacji morfosyntaktycznych przez tager

Załóżmy, że zarówno złoty standard jak i tager dopuszczają kilka znaczników. Rozpatrzmy następujące przypadki:

s	**t**	**ZS**	**T1**	**T2**	**T3**
pijanego	gen	0	0	1	1
	acc	1	1	0	1

s	**t**	**ZS**	**T1**	**T2**	**T3**	**T4**
pijaną	ppas	0	0	0	1	1
	adj:acc	1	1	0	1	1
	adj:inst	1	1	1	1	0

Przypisywanie segmentom intepretacji morfosyntaktycznych przez tager

Możliwe przypisanie interpretacji przez tager *Ti* :

- $I_{ZS}(s) = I_{Ti}(s)$
- $I_{ZS}(s) \subseteq I_{Ti}(s)$
- $I_{ZS}(s) \supseteq I_{Ti}(s)$
- $I_{ZS}(s) \cap I_{Ti}(s) \neq \emptyset$
- $I_{ZS}(s) \cap I_{Ti}(s) = \emptyset$

Poprawna intepretacja

Oczywiście chcemy, aby nasz tager dla każdego segmentu s zachowywał się tak, że $I_{ZS}(s) = I_{Ti}(s)$.

Co z pozostałymi sytuacjami? Czy możemy je czasem uznać za wystarczająco dobre ?

- tager jest bazą dla parsera składniowego ($I_{ZS}(s) \supseteq I_{Ti}(s)$)
- tager regułowy jest bazą dla tagera stochastycznego ($I_{ZS}(s) \subseteq I_{Ti}(s)$)

Poprawna intepretacja

Oczywiście chcemy, aby nasz tager dla każdego segmentu s zachowywał się tak, że $I_{ZS}(s) = I_{Ti}(s)$.

Co z pozostałymi sytuacjami? Czy możemy je czasem uznać za wystarczająco dobre ?

- tager jest bazą dla parsera składniowego ($I_{ZS}(s) \supseteq I_{Ti}(s)$)
- tager regułowy jest bazą dla tagera stochastycznego ($I_{ZS}(s) \subseteq I_{Ti}(s)$)

Poprawna intepretacja

Oczywiście chcemy, aby nasz tager dla każdego segmentu s zachowywał się tak, że $I_{ZS}(s) = I_{Ti}(s)$.

Co z pozostałymi sytuacjami? Czy możemy je czasem uznać za wystarczająco dobre ?

- tager jest bazą dla parsera składniowego ($I_{ZS}(s) \supseteq I_{Ti}(s)$)
- tager regułowy jest bazą dla tagera stochastycznego ($I_{ZS}(s) \subseteq I_{Ti}(s)$)

Definicje miar

- **Poprawność** (*ang. correctness*)

$$c(Ti, S_s) = \frac{|\{s \in S_s : I_{ZS}(s) = I_{Ti}(s)\}|}{|S_s|} \cdot 100\%$$

- **Trafność** (*ang. accuracy*)

$$acc(Ti, S_{(s,t)}) = \frac{|\{(s,t) \in S_{(s,t)} : i_{ZS}(s,t) = i_{Ti}(s,t)\}|}{|S_{(s,t)}|} \cdot 100\%$$

- **Dokładność** (*ang. precision*)

$$P(Ti, S_{(s,t)}) = \frac{|\{(s,t) \in S_{(s,t)} : i_{ZS}(s,t) = 1 \wedge i_{Ti}(s,t) = 1\}|}{|\{(s,t) \in S_{(s,t)} : i_{Ti}(s,t) = 1\}|}$$

Definicje miar (2)

- **Pełność** (*ang. recall*)

$$R(Ti, S_{(s,t)}) = \frac{|\{(s,t) \in S_{(s,t)} : i_{ZS}(s,t) = 1 \wedge i_{Ti}(s,t) = 1\}|}{|\{(s,t) \in S_{(s,t)} : i_{ZS}(s,t) = 1\}|}$$

- **F**

$$F(\alpha, Ti, S_{(s,t)}) = \frac{1}{\alpha \frac{1}{P(Ti,S_{(s,t)})} + (1-\alpha)\frac{1}{R(Ti,S_{(s,t)})}}$$

$$F(2, Ti, S_{(s,t)}) = \frac{2P(Ti, S_{(s,t)})R(Ti, S_{(s,t)})}{P(Ti, S_{(s,t)}) + R(Ti, S_{(s,t)})}$$

Najczęściej stosowane miary

Miary najczęściej stosowane przy ewaluacji tagerów dopuszczających niejednoznaczności:

- Przy ewaluacji korpusu Penn Treebank Brown (dopuszcza niejednoznaczności) stosowano miarę poprawność.
- Do ewaluacji tagerów wykorzystywanych w korpusie IPI PAN stosowano najczęściej miarę trafność.

Przykład 1

s	t	ZS	T1	T2
s1	z1	1	1	1
	z2	1	0	1
	z3	0	0	0
s2	x1	0	0	1
	x2	1	1	0
	x3	1	0	0
s3	y1	0	0	1
	y2	1	1	0
	Y3	1	0	0

$$c(T1, S_s) = 0\% \ < \ c(T2, S_s) = 33,(3)\%$$

$$acc(T1, S_{(s,t)}) = 66,(6)\% \ > \ acc(T2, S_{(s,t)}) = 33,(3)\%$$

Przykład 2

s	**t**	**ZS**	**T1**	**T2**	liczba
s1	z1	1	0	1	2
	z2	0	1	0	
	z3	0	1	0	
	z4	0	1	0	
s2	x1	0	0	1	3
	x2	1	1	0	

$$c(T1, S_s) = 60\% \ > \ c(T2, S_s) = 40\%$$

$$acc(T1, S_{(s,t)}) = 6/14 \cdot 100\% \ < \ acc(T2, S_{(s,t)}) = 8/14 \cdot 100\%$$

Przykład 1 (jeszcze raz)

s	t	ZS	T1	T2
s1	z1	1	1	1
	z2	1	0	1
	z3	0	0	0
s2	x1	0	0	1
	x2	1	1	0
	x3	1	0	0
s3	y1	0	0	1
	y2	1	1	0
	Y3	1	0	0

$$P(T1, S_{(s,t)}) = 1 \ > \ P(T2, S_{(s,t)}) = \frac{1}{2}$$
$$R(T1, S_{(s,t)}) = \frac{1}{2} \ > \ R(T2, S_{(s,t)}) = \frac{1}{3}$$
$$F(1/2, T1, S_{(s,t)}) = \frac{2}{3} \ > \ F(1/2, T2, S_{(s,t)}) = \frac{2}{5}$$

Przykład 2 (jeszcze raz)

s	**t**	**ZS**	**T1**	**T2**	liczba
s1	z1	1	0	1	2
	z2	0	1	0	
	z3	0	1	0	
	z4	0	1	0	
s2	x1	0	0	1	3
	x2	1	1	0	

$$P(T1, S_{(s,t)}) = \frac{1}{3} \; < \; P(T2, S_{(s,t)}) = \frac{2}{5}$$
$$R(T1, S_{(s,t)}) = \frac{3}{5} \; > \; R(T2, S_{(s,t)}) = \frac{2}{5}$$
$$F(1/2, T1, S_{(s,t)}) = \frac{3}{7} \; > \; F(1/2, T2, S_{(s,t)}) = \frac{2}{5}$$

Przykład 3a

s	t	ZS	T1	T2
s1	z1	1	0	1
	z2	1	1	1
	z3	0	1	0
s2	x1	0	0	1
	x2	1	1	0

$c(T1, S_s) = c(T2, S_s) = 50\%$
$acc(T1, S_{(s,t)}) = acc(T2, S_{(s,t)}) = 60\%$
$P(T1, S_{(s,t)}) = P(T2, S_{(s,t)}) = \frac{2}{3}$
$R(T1, S_{(s,t)}) = R(T2, S_{(s,t)}) = \frac{2}{3}$
$F(1/2, T1, S_{(s,t)}) = F(1/2, T2, S_{(s,t)}) = \frac{2}{3}$

Przykład 3b

s	**t**	**ZS**	**T1**	**T2**	liczba
s1	z1	1	0	1	10
	z2	1	1	1	
	z3	0	1	0	
s2	x1	0	0	0	80
	x2	1	1	1	
s3	x1	0	0	1	10
	x2	1	1	0	

$$c(T1, S_s) = c(T2, S_s) = 90\%$$
$$acc(T1, S_{(s,t)}) = acc(T2, S_{(s,t)}) = \frac{190}{210} \cdot 100\%$$
$$P(T1, S_{(s,t)}) = P(T2, S_{(s,t)}) = \frac{100}{110}$$
$$R(T1, S_{(s,t)}) = R(T2, S_{(s,t)}) = \frac{100}{110}$$
$$F(1/2, T1, S_{(s,t)}) = F(1/2, T2, S_{(s,t)}) = \frac{100}{110}$$

Spostrzeżenia

- Przykłady 1 i 2 pokazują, że miary trafność i poprawność mogą dawać mylące wyniki.
- Na trafność duży wpływ mają poprawnie rozpoznane błędne interpretacje.
- Dokładność i pełność są czułe na małe zmiany poprawnie rozpoznanych jako dobre i błednie rozpoznanych interpretacji.
- Sterując parametrem α możemy decydować czy bardziej zależy nam na dokładności czy na pełności.
- Przykłady 3a i 3b pokazują, że zdefiniowane miary okazują się czasami niewystarczające.

Wnioski

Czy fakt, że zdefiniowane miary okazują się niewystarczające, stanowi problem?

Tak: jeśli segmentów, dla których dopuszczamy więcej niż jedną poprawną interpretację, jest istotnie dużo.

Wnioski

Czy fakt, że zdefiniowane miary okazują się niewystarczające, stanowi problem?

Tak: jeśli segmentów, dla których dopuszczamy więcej niż jedną poprawną interpretację, jest istotnie dużo.

Dokładność i pełność segmentów

- **Dokładność segmentu**

$$P(Ti, s) = \frac{|I_{Ti}(s) \cap I_{ZS}(s)|}{|I_{Ti}(s)|}$$

- **Pełność segmentu**

$$R(Ti, s) = \frac{|I_{Ti}(s) \cap I_{ZS}(s)|}{|I_{ZS}(s)|}$$

Przykład 3b (jeszcze raz)

s	**t**	**ZS**	**T1**	**T2**	liczba
s1	z1	1	0	1	10
	z2	1	1	1	
	z3	0	1	0	
s2	x1	0	0	0	80
	x2	1	1	1	
s3	x1	0	0	1	10
	x2	1	1	0	

	$P(T1, s)$	$P(T2, s)$	$R(T1, s)$	$R(T2, s)$
s1	1/2	1	1/2	1
s2	1	1	1	1
s3	1	0	1	0
$E(-)$	95/100	90/100	95/100	90/100

Nowy sposób analizowania wyników

Potraktujmy $P(Ti, -)$ i $R(Ti, -)$ jako zmienne losowe. Możemy wtedy m.in. policzyć:

- jaki procent segmentów ma $R(Ti, s) > 0.5$
- $E(P(Ti, s))$, $E(R(Ti, s))$ oraz $Var(P(Ti, s))$, $Var(P(Ti, s))$
- Porównywać dystrybuanty (np.: metodą QQ plot)
- Porównywać interesujące nas warunkowe zmienne losowe

Czemu policzenie dokładności i pełności segmentów może być interesujące?

- Na miary dokładność i pełność duży wpływ mają segmenty, dla których dopuszczamy więcej niż jedną poprawną iterpretację
- Miary pełność i dokładność segmentu nie są czułe na segmenty z wieloma poprawnymi interpretacjami
- Kiedy tworzymy nowe narzędzie korzystające z tagera chcielibyśmy wiedzieć np. czy błędy są „skupione” ($Var(P(Ti, s))$ większa)
- Pozwalają odpowiedzieć na pytanie czy średnio $I_{ZS}(s) \supseteq I_{Ti}(s)$?

Wyniki (całość)

Plik etc:

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
R_D	11	0.8439	0.0215	0.8381	0.8116	0.8853
P_D	11	0.8836	0.0124	0.8777	0.8662	0.9020
F_D	11	0.8631	0.0135	0.8604	0.8461	0.8936
R_T	11	0.8573	0.0203	0.8550	0.8315	0.8993
P_T	11	0.8971	0.0118	0.8968	0.8791	0.9147
F_T	11	0.8766	0.0108	0.8749	0.8596	0.8963

Plik frek:

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
R_D	16	0.8840	0.0486	0.9175	0.8035	0.9307
P_D	16	0.8851	0.0500	0.8909	0.7917	0.9390
F_D	16	0.8822	0.0172	0.8853	0.8524	0.9061
R_T	16	0.8525	0.0415	0.8661	0.7868	0.9044
P_T	16	0.9029	0.0210	0.9014	0.8687	0.9331
F_T	16	0.8760	0.0164	0.8705	0.8533	0.9012

Wyniki (bez znaków interpunkcyjnych)

Plik etc:

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
R_D	9	0.8140	0.0275	0.8078	0.7773	0.8623
P_D	9	0.8596	0.0165	0.8539	0.8382	0.8819
F_D	9	0.8359	0.0176	0.8314	0.8112	0.8720
R_T	9	0.8308	0.0265	0.8313	0.7945	0.8795
P_T	9	0.8757	0.0128	0.8726	0.8625	0.8974
F_T	9	0.8524	0.0136	0.8516	0.8274	0.8739

Plik frek:

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
R_D	14	0.8762	0.0539	0.9078	0.7696	0.9212
P_D	14	0.8614	0.0523	0.8696	0.7703	0.9256
F_D	14	0.8661	0.0200	0.8682	0.8342	0.8918
R_T	14	0.8376	0.0486	0.8570	0.7484	0.8917
P_T	14	0.8826	0.0184	0.8836	0.8512	0.9091
F_T	14	0.8585	0.0222	0.8564	0.8210	0.8869

Wyniki (POS)

Plik etc:

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
R_D	11	0.9626	0.0071	0.9654	0.9500	0.9701
P_D	11	0.9581	0.0075	0.9599	0.9471	0.9674
F_D	11	0.9603	0.0073	0.9631	0.9486	0.9687
R_T	11	0.9505	0.0102	0.9540	0.9329	0.9648
P_T	11	0.9315	0.0117	0.9331	0.9155	0.9481
F_T	11	0.9409	0.0104	0.9420	0.9246	0.9564

Plik frek:

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
R_D	16	0.9806	0.0041	0.9807	0.9716	0.9867
P_D	16	0.9765	0.0046	0.9773	0.9684	0.9825
F_D	16	0.9785	0.0043	0.9790	0.9700	0.9846
R_T	16	0.9626	0.0124	0.9656	0.9368	0.9780
P_T	16	0.9434	0.0154	0.9456	0.9137	0.9638
F_T	16	0.9529	0.0129	0.9565	0.9251	0.9676

Uwagi

- Aby tagsety stosowane przez tagery i złoty standard był takie same rzutowaliśmy je
- Przy obliczeniach uwzględniamy tylko te segmenty, które są takie same wg. złotego standardu i danego tagera
- Obliczenia zostaną powtórzone, gdy na nowo wytrenuje się tagery

KONIEC

DZIĘKUJĘ ZA UWAGĘ