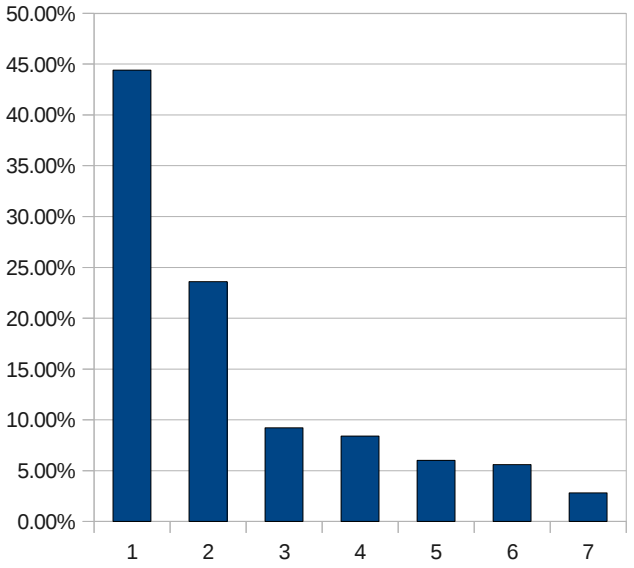
Míra věrohodné informace v anotovaných datech a Regulace granularity sémantických kategorií

Martin Holub a Vincent Kríž

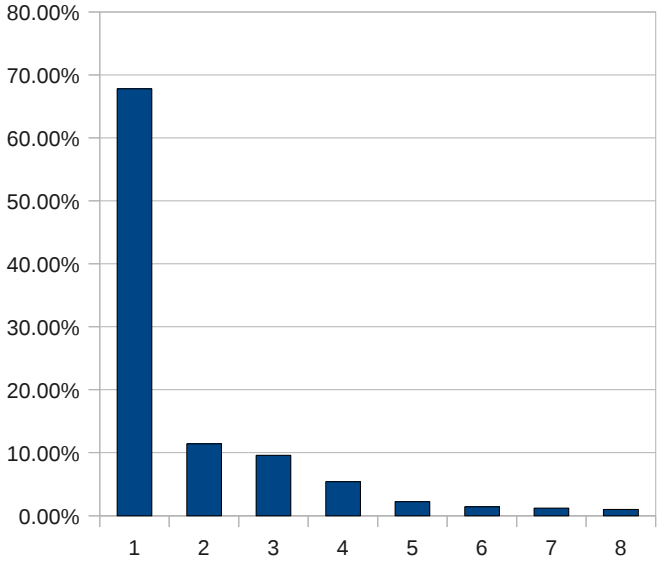
Obtížnost rozhodnutí o přiřazení sémantické kategorie z pohledu teorie informace

- Počet přiřazovaných značek vs. perplexita jejich rozdělení
- větší perplexita → správná značka sděluje *více informace*

WAKE
větší perplexita



CLAIM
menší perplexita



Verb	1	2	3	4	5	6	7	8	Total
claim	27	339	57	48	11	5	6	7	500
wake	23	7	14	111	59	15	21	0	250

Optimální granularita sémantických kategorií?

Správné řešení:
optimalizovat vzhledem k definované aplikaci

- čím jemněji jsou sémantické kategorie rozlišené, tím detailnější informaci aplikace získává (pokud je přiřazení sémantické kategorie správné)

- ALE PROTI TOMU: větší granularita implikuje větší míru nejistoty při disambiguaci (jak manuální, tak automatické)

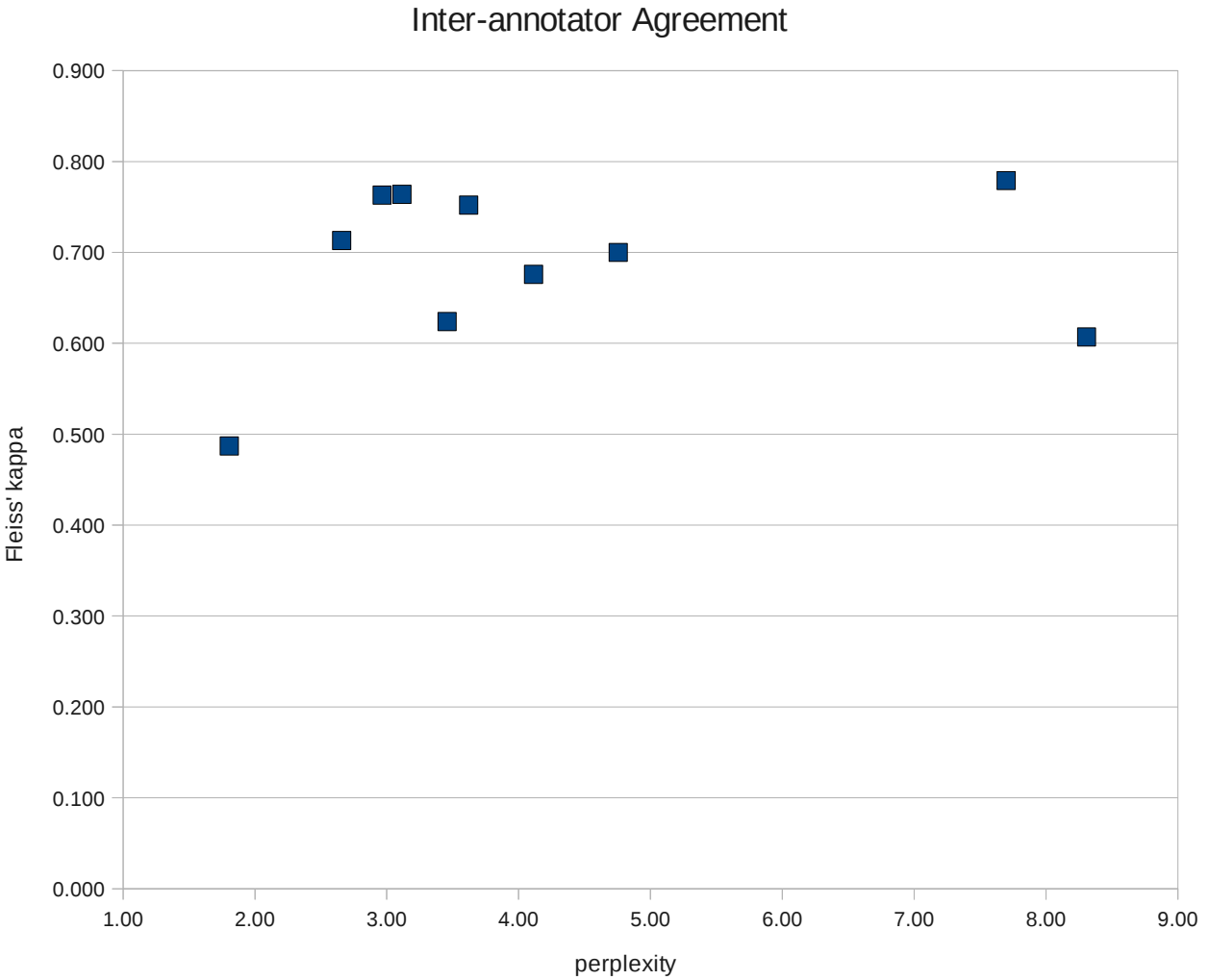
- nemá smysl vytvářet zbytečně sémantické kategorie, jejichž rozlišování by bylo pro danou aplikaci irelevantní

Granularita sémantických kategorií bez určené aplikace?

- Manuální anotace spolehlivě ukazují, že přílišná granularita vede k vysoké míře mezianotátorské neshody, protože definované rozdíly mezi sémantickými kategoriemi jsou příliš vágní/jemné
- Kde je "rozumná mez" pro maximum granularity, pokud cílová aplikace není určena?
  - tam, kde končí schopnost lidí shodnout se na určení správné sémantické kategorie
  - jistou míru mezianotátorské neshody je nutno připustit – vyloučit ji zcela nelze
  - vždy však chceme, aby informační zisk byl (v průměrném případě) vyšší než ztráta způsobená nejistotou o správnosti sémantické kategorie

Pohled na anotovaná data optikou míry mezianotátorské shody

Verbs	Perplexity
ally	4.11
arrive	2.96
claim	3.12
cry	3.62
enlist	3.46
halt	1.81
plough	7.70
plug	8.31
submit	2.66
wake	4.75



Pohled na anotovaná data optikou matic konfuze

- Matice konfuze
  - 3 anotátoři
  - 5 různých značek
  - 50 anotovaných instancí

  sloveso *submit*

A_1 vs. A_2					
	1	1.a	2	4	5
1	29	1	1	0	0
1.a	0	1	0	0	0
2	0	1	11	0	0
4	0	0	0	2	0
5	0	0	0	3	1

A_1 vs. A_3					
	1	1.a	2	4	5
1	29	2	0	0	0
1.a	1	0	0	0	0
2	0	0	12	0	0
4	0	0	0	1	1
5	0	0	0	0	4

A_2 vs. A_3					
	1	1.a	2	4	5
1	27	2	0	0	0
1.a	2	0	1	0	0
2	1	0	11	0	0
4	0	0	0	1	4
5	0	0	0	0	1

Anotační úloha a anotovaná data – formální model

- anotovaný vzorek $\mathcal{S} = \{s_1, \ldots, s_r\}$
- množina anotátorů $\mathcal{A} = \{A_1, \ldots, A_m\}$
- množina použitých značek $\mathcal{T} = \{t_1, \ldots, t_n\}$

- anotátor je funkce $A_i(s) = \{t\} \qquad s \in \mathcal{S}, t \in \mathcal{T}$

- dvojice anotátorů dává množinu

$$\{t, t'\} = A_i(s) \cup A_j(s)$$

Agregovaná matice konfuze C*

- C* je symetrická matice
- na diagonále: kolikrát se dvojice anotátorů shodla na dané značce
- mimo diagonálu: kolikrát se dvojice anotátorů NEshodla daným způsobem

	1	1.a	2	4	5
1	85	8	2	0	0
1.a	8	1	2	0	0
2	2	2	34	0	0
4	0	0	0	4	8
5	0	0	0	8	6

Pravděpodobnost užití značky jedním nebo dvěma anotátory

- pravděpodobnost, že anotátor použije značku

$$p_1(t_i) = \frac{1}{mr} \sum_{k=1}^{m} \sum_{j=1}^{r} |A_k(s_j) \cap \{t_i\}|$$

- pravděpodobnost, že dvojice anotátorů použije značku

$$p_2(t_i) = \frac{1}{\binom{m}{2} r} \sum_{k} C^{\star}_{ik}$$

- pravděpodobnost, že anotátor použije značku za předpokladu, že jiný anotátor použil nějakou danou značku

$$p_2(t_i \mid t_j) = \frac{C^{\star}_{ij}}{\binom{m}{2} r \cdot p_2(t_j)} = \frac{C^{\star}_{ij}}{\sum_k C^{\star}_{jk}}$$

Matice pravděpodobnosti konfuze C^p

Definition: *Confusion Probability Matrix* (CPM)

$$C^p_{ji} = p_2(t_i \mid t_j) = \frac{C^\star_{ij}}{\sum_k C^\star_{jk}}.$$

	1	1.a	2	4	5
1	0.895	0.084	0.021	0.000	0.000
1.a	0.727	0.091	0.182	0.000	0.000
2	0.053	0.053	0.895	0.000	0.000
4	0.000	0.000	0.000	0.333	0.667
5	0.000	0.000	0.000	0.571	0.429

Kolik *věrohodné* informace poskytuje jedna přiřazená značka?

- Tradiční míra pro množství informace $I(t_j) = -\log p_1(t_j)$
  - to lze brát vážně, pouze pokud máme shodu anotátorů
  - jestliže se anotátoři neshodují, množství informace je redukované
- Porovnejme rozdělení $p_1(t_i)$ a $p_2(t_i \mid t_j),\ i = 1, \ldots, n$
- Je-li značka "spolehlivá", pak by mělo platit

$$p_2(t_j \mid t_j) > p_1(t_j)$$

$$p_2(t_i \mid t_j) < p_1(t_i)$$

pro všechna *i* různá od *j*.

- Při stoprocentní shodě anotátorů platí $p_2(t_j \mid t_j) = 100\,\%$
- **Pro posouzení množství věrohodné informace, kterou nese jedna značka, potřebujeme tedy funkci, která porovná dvě uvedená pravděpodobnostní rozdělení a řekne, do jaké míry jsou splněny požadované vlastnosti.**

Věrohodný (informační) zisk z jedné přiřazené značky

- Formule se podobá tradičnímu *information gain*

Definition: *Reliable Gain* (RG) from the tag t_j is

$$RG(t_j) = \sum_k -(-1)^{\delta_{kj}} p_2(t_k|t_j) \log \frac{p_2(t_k|t_j)}{p_1(t_k)}.$$

- Při dokonalé shodě je maximum $RG(t_j) = -\log p_1(t_j)$
- Hodnota může RG být i negativní – pak je daná značka “kontraproduktivní”, tj. její užití nepřináší věrohodnou, ale spíše matoucí informaci.

Průměrný věrohodný zisk z přiřazené značky (= informační hodnota tagsetu)

- Definujeme jako očekávanou hodnotu RG

Definition: *Average Reliable Gain* (ARG) from the tagset $\{t_1, \ldots, t_n\}$ is computed as an expected value of $RG(t_j)$:

$$ARG = \sum_j p_1(t_j) RG(t_j)$$

- Maximální hodnota ARG – při absolutní shodě všech anotátorů – je rovna entropii rozdělení p_1

$$ARG_{max} = H(p_1(t_1), \ldots, p_1(t_n))$$

Porovnání tradiční míry pro mezianotátorskou shodu a ARG

- ARG neměří shodu, ale *množství informace* "s ohledem na shodu"
- S rostoucím IAA (e.g. Fleiss' kappa) může ARG růst nebo klesat
- Příklad:

COOL	tagset	ARG	Fleiss' kappa
orig:	1, 2, 2.a, 2.f, 4, 4.a, 6, 7, 7.a, 10, 11, 13, 17, x	**1.556**	**0.843**
merged:	2 + 11	**1.875**	**0.888**

COOL	tagset	ARG	Fleiss' kappa
orig:	1, 2, 2.a, 2.f, 4, 4.a, 6, 7, 7.a, 10, 11, 13, 17, x	**1.556**	**0.843**
merged:	1 + 13	**1.086**	**0.854**

Kdy se vyplatí dvě různé sémantické kategorie sloučit?

- Hledáme kompromis mezi mírou mezianotátorské shody a mírou granularity
- Optimalizujeme ARG (aproximace hladovým algoritmem)

ACCESS	tagset	ARG	Fleiss' kappa
orig:	1, 1.a, 1.s, 2, 2.a, 3, 3.f, 4, 4.a, 5.a, 6, u, x	**0.421**	**0.600**
optimized:	[1 + 1.a + 1.s], [2 + 2.a], [3 + 4 + 4.a + 6], [3.f], [u + 5.a], [x]	**0.979**	**0.727**

verbs	ORIG. TAGSET		REDUCED TAGSET	
	size	ARG	size	ARG
access-ref.txt	13	0.421	6	0.979
ally-ref.txt	14	0.799	7	1.382
arrive-ref.txt	11	0.779	6	1.516
breathe-ref.txt	15	1.108	6	1.724
claim-ref.txt	13	0.945	5	1.511
cool-ref.txt	14	1.556	7	2.284
crush-ref.txt	19	-0.093	4	1.408
cry-ref.txt	18	0.748	9	1.55
enlist-ref.txt	13	0.179	4	1.685
halt-ref.txt	11	0.159	3	0.495
part-ref.txt	19	1.572	10	2.771
plough-ref.txt	22	1.316	11	2.869
plug-ref.txt	27	0.384	8	1.823
submit-ref.txt	9	0.433	3	1.266
wake-ref.txt	14	0.775	6	1.413
yield-ref.txt	23	0.422	6	1.467