Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Alexander Panchenko

FROM UNSUPERVISED INDUCTION OF LINGUISTIC STRUCTURES FROM TEXT TOWARDS APPLICATIONS IN DEEP LEARNING

Shared task on word sense induction

A shared task on WSI

- An **ACL SIGSLAV** sponsored shared task on **word sense induction** (WSI) for the Russian language.
- **More details**: `http://russe.nlpub.org/2018/wsi`






A lexical sample WSI task

- **Target word**, e.g. “bank”.

A lexical sample WSI task

- **Target word**, e.g. “bank”.
- **Contexts** where the word occurs, e.g.:
  - “river **bank** is a slope beside a body of water”
  - “**bank** is a financial institution that accepts deposits”
  - “Oh, the **bank** was robbed. They took about a million dollars.”
  - “**bank** of Elbe is a good and popular hangout spot complete with good food and fun”

A lexical sample WSI task

- **Target word**, e.g. “bank”.
- **Contexts** where the word occurs, e.g.:
  - “river **bank** is a slope beside a body of water”
  - “**bank** is a financial institution that accepts deposits”
  - “Oh, the **bank** was robbed. They took about a million dollars.”
  - “**bank** of Elbe is a good and popular hangout spot complete with good food and fun”
- You need to **group the contexts by senses**:
  - “river **bank** is a slope beside a body of water”
  - “**bank** of Elbe is a good and popular hangout spot complete with good food and fun”
  - “**bank** is a financial institution that accepts deposits”
  - “Oh, the **bank** was robbed. They took about a million dollars.”

Dataset based on Wikipedia

Dataset	Type	Inventory	Corpus	Split	# of words	# of senses	Avg. # of senses	# of contexts
wiki-wiki	main	Wikipedia	Wikipedia	train	4	8	2.0	439
wiki-wiki	main	Wikipedia	Wikipedia	test	5	12	2.4	539
bts-rnc	main	BTS	RNC	train	30	96	3.2	3491
bts-rnc	main	BTS	RNC	test	51	153	3.0	6556
active-dict	main	Active Dict.	Active Dict.	train	85	312	3.7	2073
active-dict	main	Active Dict.	Active Dict.	test	168	555	3.3	3729
active-rnc	additional	Active Dict.	RNC	train	20	71	3.6	1829
active-rutenten	additional	Active Dict.	ruTenTen[12]	train	21	71	3.4	3671
bts-rutenten	additional	BTS	ruTenTen	train	11	25	2.3	956

Dataset based on RNC

Dataset	Type	Inventory	Corpus	Split	# of words	# of senses	Avg. # of senses	# of contexts
wiki-wiki	main	Wikipedia	Wikipedia	train	4	8	2.0	439
wiki-wiki	main	Wikipedia	Wikipedia	test	5	12	2.4	539
bts-rnc	main	BTS	RNC	train	30	96	3.2	3491
bts-rnc	main	BTS	RNC	test	51	153	3.0	6556
active-dict	main	Active Dict.	Active Dict.	train	85	312	3.7	2073
active-dict	main	Active Dict.	Active Dict.	test	168	555	3.3	3729
active-rnc	additional	Active Dict.	RNC	train	20	71	3.6	1829
active-rutenten	additional	Active Dict.	ruTenTen[12]	train	21	71	3.4	3671
bts-rutenten	additional	BTS	ruTenTen	train	11	25	2.3	956

Dataset based on dictionary glosses

Dataset	Type	Inventory	Corpus	Split	# of words	# of senses	Avg. # of senses	# of contexts
wiki-wiki	main	Wikipedia	Wikipedia	train	4	8	2.0	439
wiki-wiki	main	Wikipedia	Wikipedia	test	5	12	2.4	539
bts-rnc	main	BTS	RNC	train	30	96	3.2	3491
bts-rnc	main	BTS	RNC	test	51	153	3.0	6556
active-dict	main	Active Dict.	Active Dict.	train	85	312	3.7	2073
active-dict	main	Active Dict.	Active Dict.	test	168	555	3.3	3729
active-rnc	additional	Active Dict.	RNC	train	20	71	3.6	1829
active-rutenten	additional	Active Dict.	ruTenTen[12]	train	21	71	3.4	3671
bts-rutenten	additional	BTS	ruTenTen	train	11	25	2.3	956

A sample from the *wiki-wiki* dataset

word	articles	sense
белка	кавказская белка; обыкновенная белка; японская белка; капская земляная белка; аризонская белка; ...	рыжая, шустрая, дерево, вскарабкаться, спрыгнуть
белка	домен белка; биосинтез белка; фолдинг белка; институт белка ран; сигнальная функция белка; ...	желток, пища, углевод, рацион, жир
белка	белка и стрелка; белка и стрелка (мюзикл); белка и стрелка. лунные приключения; ...	космос, полет, животные, первые, советские

A sample from the *wiki-wiki* dataset

Rank	Team	ARI (public)	ARI (private)
1	jamsic	1.0000 (1)	0.9625 (1)
2	akutuzov (Kutuzov, 2018)	0.9823 (2)	0.7096 (2)
3	ezhick179 (Arefyev et al., 2018)	1.0000 (1)	0.6586 (3)
-	*akapustin*	0.6520 (6)	0.6459 (4)
-	*aby2s*	1.0000 (1)	0.5889 (5)
-	*bokan*	0.7587 (5)	0.5530 (6)
*	**AdaGram (Bartunov et al, 2016)**	**0.6278 (7)**	**0.5275 (7)**
4	Pavel (Arefyev et al., 2018)	0.9649 (3)	0.4827 (8)
5	eugenys	0.0115 (12)	0.4377 (9)
6	mikhal	1.0000 (1)	0.4109 (10)
7	fogside	0.6520 (6)	0.3958 (11)

A sample from the *wiki-wiki* dataset

Rank	Team	ARI (public)	ARI (private)
1	embeddings + sense induction		0.9625 (1)
2	embeddings + affinity propagation		0.7096 (2)
3	embeddings + agglomerative clustering		0.6586 (3)
-	*akapustin*	0.6520 (6)	0.6459 (4)
-	*aby2s*	1.0000 (1)	0.5889 (5)
-	*bokan*	0.7587 (5)	0.5530 (6)
*	**AdaGram (Bartunov et al, 2016)**	**0.6278 (7)**	**0.5275 (7)**
4	Pavel (Arefyev et al., 2018)	0.9649 (3)	0.4827 (8)
5	eugenys	0.0115 (12)	0.4377 (9)
6	mikhal	1.0000 (1)	0.4109 (10)
7	fogside	0.6520 (6)	0.3958 (11)

A sample from the *bts-rnc* dataset

Rank	Team	ARI (public)	ARI (private)
1	jamsic	0.3508 (1)	0.3384 (1)
2	Pavel (Arefyev et al., 2018)	0.2812 (2)	0.2818 (2)
-	*joystick*	0.2477 (5)	0.2579 (3)
-	*Timon*	0.2360 (7)	0.2434 (4)
3	akutuzov (Kutuzov, 2018)	0.2448 (6)	0.2415 (5)
4	ezhick179 (Arefyev et al., 2018)	0.2599 (4)	0.2284 (6)
-	*thebestdeeplearningspecialist*	0.2178 (8)	0.2227 (7)
5	fogside	0.1661 (10)	0.2154 (8)
*	**AdaGram (Bartunov et al., 2016)**	**0.2624 (3)**	**0.2132 (9)**
-	*aby2s*	0.1722 (9)	0.2102 (10)
6	bokan	0.1363 (11)	0.1515 (11)

A sample from the *active-dict* dataset

Rank	Team	ARI (public)	ARI (private)
1	jamsic	0.2643 (1)	0.2477 (1)
2	Pavel (Arefyev et al., 2018)	0.2361 (4)	0.2270 (2)
-	*Timon*	0.2324 (5)	0.2222 (3)
-	*thebestdeeplearningspecialist*	0.2297 (6)	0.2194 (4)
3	akutuzov (Kutuzov, 2018)	0.2396 (3)	0.2144 (5)
-	*aby2s*	0.2465 (2)	0.1985 (6)
-	*joystick*	0.1890 (8)	0.1939 (7)
4	ezhick179 (Arefyev et al., 2018)	0.1899 (7)	0.1839 (8)
*	**AdaGram (Bartunov et al., 2016)**	**0.1764 (9)**	**0.1538 (9)**
-	*ostruyanskiy*	0.1515 (10)	0.1403 (10)
-	*akapustin*	0.1337 (11)	0.1183 (11)

jamsic: sense induction

1. **Get the neighbors** of a target word, e.g. **"bank"**:
   1. lender
   2. river
   3. citybank
   4. slope
   5. ...
2. Get **similar to "bank"** and **dissimilar to "lender"**:
   1. river
   2. slope
   3. land
   4. ...
3. **Compute distances** to **"lender"** and **"river"**.