A Appendices

A.1 Details of the parameters used to run the systems

The parameters that are used to run NEO system:

```
-targ aa \
-corpus_type 2 \
-files <REDACTED> \
-whether_csv_suffix no \
-group_size 1 \
-left 5 \
-right 5 \
-whether_pad no \
-lowercase no \
-corp_wc_min 0 \
-whether_archive no \
-expts <REDACTED> \
-num_senses 2 \
-unsup yes \
-dynamic yes \
-senseprobs_string .500/.500 \
-word_rand yes \
-words_prior_type corpus \
-rand_word_mix yes \
-rand_word_mix_level 0.00001 \
-save_starts yes \
-time_stamp yes \
-max_it 500 \
-record_version yes

-targ aa
-corpus_type 2
-files <REDACTED>
-whether_csv_suffix no
-group_size 1
-left 5
-right 5
-whether_pad no
-lowercase no
-filter_ngram_meta_token no
-syn_info 0
 excl_targ_nnp is 1
 excl_targ_np1 is 1
 note both above set by -excl_targ_name yes/no
-wlp 1
-targ_wlp 0
-filter_coha_at_token no

-expts <REDACTED>
-num_senses 2
-unsup yes
-sup no
-dynamic yes
-senseprobs_string .500/.500
-sense_rand no
-rand_mix no
-words_prior_type corpus
-word_rand yes
-rand_word_mix yes
-rand_word_mix_level 1e-05
-set_seed yes
-max_it 500
-wordprobs_out em_wordprobs_final
-senseprobs_out em_senseprobs_final
-save_starts yes
-time_stamp yes
-record_version yes
-whether_save_cat_outcomes no
-whether_save_word_stats no
```

The parameters which are used to run SCAN system:

```
text_corpus <REDACTED>
target_words <REDACTED>
bin_corpus_store <REDACTED>
window_size 5
full_corpus_path <REDACTED>
word_corpus_path <REDACTED>
output_path <REDACTED>
kappaF 10.0
kappaK 4.0
a0 7.0
b0 3.0
num_top 2
iterations 500
start_time 1945
end_time 2018
time_interval 1
min_doc_per_word 0
max_docs_per_slice 9999999999
```

The parameters which are given to the emergence algorithm “EmergeTime”:

```
ThresholdValue 0
WindowSize 5
Step_increase_threshold_wrtMax 0.04
Min_total_surges_window_increase_wrtMax 0.2
Max_threshold_previouslowyears_wrtMax 0.1
Min_prop_previouslowyears_wrtoutsidetheWindow 0.8
```

A.2 Dataset Summary

- Medline version: 2019.
- Pubmed Central version: 2019.
- UMLS version: 2018AB-full.
- Language: English.

Table 10 shows the gold standard output (senses and year of emergence), as obtained by the "EmergeTime" emergence detection algorithm based on the original gold data.

Table 11 shows the frequency of ocurrences of each Concept for each sense of the ambigious word as appears in the MEDLINE citations as well as PubMed Central.

Table 12 shows a summary of the final dataset for each ambiguous word.

A.3 Matching Method Comparisons

This section provides a brief explanation about the issues identified in the two instance-based evaluation methods proposed by (Agirre and Soroa, 2007) for the SemEval 2007 Shared Task and by (Manandhar et al., 2010) for the SemEval 2010 Shared Task. Throughout this section we use the example represented in the following confusion matrix, where G_i denotes the gold standard sense i and C_j denotes the induced cluster j:

C_j	G_1	G_2	G_3
C_0	15	8	10
C_1	1	11	30
C_2	60	1	20

A.3.1 The unsupervised evaluation setting

In this setting, the perfomance is calculated based on matching every gold sense to one of the predicted clusters. This is done by calculating for every possible pair (G_i, C_j) the F1-score obtained by matching G_i with C_j, then selecting the predicted cluster with the maximum F1-score for every gold sense. By design, this matching method allows different gold classes to be assigned to the same predicted cluster. In particular this is likely to happen when there is imbalance between senses: in such a case, the majority class will influence the assessment of the quality of a clustering solution, since the final the F1-score will be selected based on the majority class. The table below shows the F1-score for every cluster C_j and every class G_i for the example presented above. It can be observed that both G_2 and G_3 are matched against cluster C_1.

C_j	G_1	G_2	G_3
C_0	0.27	0.30	0.21
C_1	0.01	**0.35**	**0.58**
C_2	**0.76**	0.01	0.28

A.3.2 The supervised evaluation setting

In this setting which is used in both (Agirre and Soroa, 2007) and (Manandhar et al., 2010), there is no direct matching between the gold standard senses and the predicted clusters. Instead a probabilistic method is used: the mapping is represented by a matrix where each cell represents the conditional probability $P(G_i|C_j)$ for every pair of predicted cluster and gold sense, and is calculated from the "training set". It is worth noticing that the same gold class G_i can obtain the highest probability $P(G_i|C_j)$ with several predicted clusters,

Target	number of senses	CUI(Sense)	Emergence Year	First Year Occurence	Target	number of senses	CUI(Sense)	Emergence Year	First Year Occurence
CCl4	2	C0209338	1994	1991	Cold	3	C0009443	1945	1945
Cold	3	C0024117	1998	1959	CCD	2	C0751951	1997	1965
GAG	2	C0017346	1988	1982	Pharmaceutical	2	C0013058	1963	1963
CAM	2	C0178551	2002	2003	Language	2	C0033348	1986	1958
EMS	2	C0015063	1974	1975	TNC	2	C0076088	1983	1985
SCD	2	C0085298	1988	1950	Synapsis	2	C0598501	1998	1951
EM	2	C0014921	1973	1975	Sodium	2	C0037570	1945	1945
Plague	2	C0032066	1959	1946	CP	3	C0033477	1971	1946
CIS	2	C0162854	1991	1992	PHA	2	C0030779	2002	1976
IP	2	C0021069	2000	1989	Leishmaniasis	2	C1548483	2005	1947
IA	2	C0021487	1946	1946	rDNA	2	C0012933	1980	1981
CDA	2	C0092801	1982	1983	HIV	2	C0019693	1987	1987
OCD	2	C0029421	1983	1984	PCA	5	C0030131	1972	1974
PCA	5	C0078944	1987	1989	PCA	5	C0149576	1957	1957
PCA	5	C0429865	1999	1960	LABOR	2	C0022864	1945	1945
CH	2	C0039021	1946	1946	Tax	2	C0144576	1992	1983
HCl	2	C0023443	1975	1954	Gas	2	C0016204	1945	1945
PEP	2	C0135981	1978	1980	TPA	2	C0032143	1983	1982
Eels	2	C0677644	2003	2004	Fe	2	C0376520	1995	1946
Cortical	3	C0001613	1945	1945	NPC	2	C0220756	2005	2006
DON	2	C0028652	1979	1981	CTX	2	C0238052	1997	1974
NEUROFIBROMATOSIS	2	C0162678	1990	1991	SARS	2	C1175743	2002	2002
TSF	2	C0021756	1976	1977	Orf	2	C0079941	1986	1982
ADP	2	C0004374	1958	1959	dC	2	C0011485	1971	1973
lens	3	C0023308	1951	1952	MCC	2	C0162804	1990	1991
SS	2	C0085077	1990	1964	WT1	2	C0148873	1991	1991
MAF	2	C0919482	2001	1998	Ice	3	C0534519	1990	1991
Lupus	3	C0024138	1945	1946	Fish	2	C0162789	1990	1953
DDS	3	C0085104	1988	1990	DDS	3	C0950121	1999	2001
drinking	2	C0684271	1946	1946	JP	2	C0031106	1946	1947
Pleuropneumonia	2	C0026934	1945	1945	NM	2	C0027972	1963	1946
NBS	2	C0398791	2003	2002	TPO	2	C0021965	1974	1975
SARS-associated coronavirus	2	C1175743	2002	2002	PR	2	C0034833	1972	1973
eCG	2	C0018064	1989	1945	MBP	2	C0065661	1999	1984
US	2	C0041618	1971	1945	FTC	2	C0206682	1992	1993
ERP	2	C0008310	1978	1978	Phosphorylase	2	C0917783	2005	1973
Ion	2	C0022024	1945	1946	Hemlock	2	C0242872	2004	2002
TLC	2	C0040509	1974	1959	Wasp	2	C0258432	1993	1994
INDO	2	C0021246	1961	1963	ADH	2	C0001942	1978	1976
Ala	3	C0002563	1954	1953	Ganglion	2	C0017067	1946	1946
Ganglion	2	C1258666	2006	1946	BPD	2	C0006287	1980	1981
Potassium	2	C0162800	1990	1948	CPDD	2	C0008838	1971	1972
THYMUS	3	C0040112	1948	1949	Malaria	2	C0206255	1991	1945
Cell	2	C1136359	2010	1999	ANA	2	C0002463	1962	1963
ORI	2	C0242961	1993	1993	cRNA	2	C0056208	1981	1982
CAD	2	C1956346	1983	1985	BSE	2	C0085209	1991	1991
Coffee	2	C0085952	2001	1962	SPR	2	C0597731	1996	1998
WBS	2	C0175702	1994	1995	Cortex	2	C0001614	1948	1950
TAT	3	C0017375	1988	1989	TAT	3	C0039341	1983	1985
Glycoside	2	C0017977	1946	1946	DAT	2	C0114838	1989	1989
Ca	3	C0006754	1945	1945	DBA	2	C1260899	1999	2001

Table 10: Goldstandard set by emergence detection algorithm “EmergeTime”. The table includes two type of emrgence: the “emergence year” which is provided by the algorithm and the “first year occurence” which indicates the first year appearance of a sense in the data

CUI(Sense)	Freq	Target	CUI(Sense)	Freq	Target	CUI(Sense)	Freq	Target	CUI(Sense)	Freq	Target
C0001972	2273	AA	C0011037	698	DDD	C0024138	6291	Lupus	C0035331	10523	Retinal
C0002520	301290	AA	C0026256	2423	DDD	C0024141	249403	Lupus	C0040452	34170	Root
C0001457	22184	ADA	C0010980	9593	DDS	C0079786	1425	MAF	C0242726	342033	Root
C0002456	1299	ADA	C0085104	56838	DDS	C0919482	1178	MAF	C1175175	33076	SARS
C0001942	20919	ADH	C0950121	326	DDS	C0014063	30365	MBP	C1175743	29333	SARS
C0003779	11012	ADH	C0011198	1684	DE	C0065661	22323	MBP	C1175175	33076	SARS-associated coronavirus
C0001459	54815	ADP	C0017480	80738	DE	C0007129	12514	MCC	C1175743	29333	SARS-associated coronavirus
C0004374	579	ADP	C0011848	9160	DI	C0162804	511	MCC	C0002895	57045	SCD
C0002736	125309	ALS	C0032246	22524	DI	C0024518	43630	MHC	C0085298	23885	SCD
C0003372	4332	ALS	C0012020	925	DON	C0027100	18485	MHC	C0037231	539	SLS
C0002463	1234	ANA	C0028652	1355	DON	C0024487	45898	MRS	C0037506	29240	SLS
C0003243	16995	ANA	C0012238	26483	Digestive	C0025235	1227	MRS	C0164209	4815	SPR
C0001625	74305	Adrenal	C0012240	9757	Digestive	C0024530	332626	Malaria	C0597731	34855	SPR
C0014563	97328	Adrenal	C0013710	42611	EGG	C0206255	25609	Malaria	C0039101	10219	SS
C0001898	38827	Ala	C0029974	57923	EGG	C0001629	10489	Medullary	C0085077	4191	SS
C0002563	29471	Ala	C0014921	2786	EM	C0025148	23198	Medullary	C0162731	1529	STEM
C0051405	13117	Ala	C0026019	92207	EM	C0026131	224599	Milk	C0242767	25397	STEM
C0225984	790	Arteriovenous Anastomoses	C0013961	37565	EMS	C0026140	81129	Milk	C0036319	12759	Schistosoma mansoni
C0684204	803	Arteriovenous Anastomoses	C0015063	4629	EMS	C0027960	12451	Moles	C0036330	8277	Schistosoma mansoni
C0039277	6282	Astragalus	C0008310	30216	ERP	C0324740	2684	Moles	C0037473	265756	Sodium
C0330845	2603	Astragalus	C0015214	54672	ERP	C0027819	80894	NBS	C0037570	33518	Sodium
C0023434	93519	B-Cell Leukemia	C0015230	10625	ERUPTION	C0398791	970	NBS	C0038160	7046	Staph
C2004493	288	B-Cell Leukemia	C1533692	7678	ERUPTION	C0085113	9357	NEUROFIBROMATOSIS	C0038170	30303	Staph
C0006298	43500	BAT	C0013671	7236	Eels	C0162678	2359	NEUROFIBROMATOSIS	C0038280	13603	Sterilization
C0008139	88303	BAT	C0677644	347	Eels	C0025033	6146	NM	C0038288	5469	Sterilization
C0005740	55247	BLM	C0014563	97328	Epi	C0027972	4093	NM	C0038395	6581	Strep
C0005859	6356	BLM	C0014582	19296	Epi	C0028587	17697	NPC	C0038402	29450	Strep
C0006012	29371	BPD	C0014772	11901	Erythrocytes	C0220756	3897	NPC	C0039062	125608	Synapsis
C0006287	21233	BPD	C0014792	353039	Erythrocytes	C0006147	141203	Nurse	C0598501	4857	Synapsis
C0006137	140353	BR	C0015259	623402	Exercises	C0028661	98472	Nurse	C0017375	2214	TAT
C0006222	8196	BR	C0452240	30855	Exercises	C0006147	139863	Nursing	C0039341	27001	TAT
C0005902	7780	BSA	C0015625	18453	FA	C0028677	32380	Nursing	C0039756	375	TAT
C0036774	65211	BSA	C0016410	66996	FA	C0028768	53452	OCD	C0040975	8	TEM
C0085105	5677	BSE	C0041713	588	FTC	C0029421	4012	OCD	C0678118	38610	TEM
C0085209	15801	BSE	C0206682	6844	FTC	C0028905	8361	OH	C0040112	965	THYMUS
C0006033	7632	Borrelia	C0032580	28787	Familial Adenomatous Polyposis	C0063146	29590	OH	C0040113	117518	THYMUS
C0024198	37268	Borrelia	C0162832	18516	Familial Adenomatous Polyposis	C0206601	2221	ORI	C1015036	1806	THYMUS
C0006304	4504	Brucella abortus	C0302583	531868	Fe	C0242961	13154	ORI	C0008569	21429	TLC
C0302363	2154	Brucella abortus	C0376520	40837	Fe	C0013570	1832	Orf	C0040509	3718	TLC
C0011905	5723	CAD	C0016163	200509	Fish	C0079941	84657	Orf	C0039493	22957	TMJ
C1956346	165919	CAD	C0162789	104590	Fish	C0033036	1149	PAC	C0039496	6062	TMJ
C0007578	31452	CAM	C0018120	45586	Follicle	C0949780	204	PAC	C0040079	924	TMP
C0178551	7568	CAM	C0221971	26046	Follicle	C0032172	55064	PAF	C0041041	14171	TMP
C0008928	2201	CCD	C0018120	103753	Follicles	C0037019	590	PAF	C0076088	20377	TNC
C0751951	950	CCD	C0221971	23601	Follicles	C0030131	2313	PCA	C0077400	6334	TNC
C0007022	19811	CCl4	C0017346	4737	GAG	C0030625	2766	PCA	C0041070	6365	TNT
C0209338	3149	CCl4	C0017973	45430	GAG	C0078944	12351	PCA	C0077404	17258	TNT
C0002876	1338	CDA	C0017067	15625	Ganglion	C0149576	2037	PCA	C0032143	52834	TPA
C0092801	6319	CDA	C1258666	2093	Ganglion	C0429865	65444	PCA	C0039654	89585	TPA
C0011485	5467	CDR	C0016204	2599	Gas	C0032447	93888	PCB	C0021965	15262	TPO
C0021024	5069	CDR	C0017110	53043	Gas	C0033223	2722	PCB	C0040052	15428	TPO
C0008115	359471	CH	C0007158	3755	Glycoside	C0022521	6989	PCD	C0021759	4254	TRF
C0039021	25247	CH	C0017977	21871	Glycoside	C0162638	1233594	PCD	C0040162	43836	TRF
C0008107	20392	CI	C0020259	10437	HCl	C0030855	8811	PCP	C0021756	177829	TSF
C0022326	6259	CI	C0023443	9878	HCl	C0031381	19265	PCP	C0040052	15633	TSF
C0007099	16439	CIS	C0021760	359798	HGF	C0031642	4608	PEP	C0041484	35723	TYR
C0162854	268	CIS	C0062534	72879	HGF	C0135981	1381	PEP	C0041485	116975	TYR
C0265252	358	CLS	C0036220	43775	HHV 8	C0030779	319	PHA	C0039371	17230	Tax
C0343084	870	CLS	C0376526	53066	HHV 8	C0031858	21092	PHA	C0144576	107373	Tax
C0007789	59567	CP	C0019682	131115	HIV	C0017360	3182	POL	C0013220	41988	Tolerance
C0008925	20764	CP	C0019693	1619519	HIV	C0032356	35172	POL	C0020963	50759	Tolerance
C0033477	6615	CP	C0079504	1925	HPS	C0034044	9655	PR	C0010414	16808	Torula
C0008838	273762	CPDD	C0242994	2350	HPS	C0034833	67210	PR	C0010415	6893	Torula
C0553730	2951	CPDD	C0010343	9044	HR	C0032624	9305	PVC	C0041618	350047	US
C0010132	75466	CRF	C0018810	281856	HR	C0151636	8629	PVC	C0041703	454576	US
C0022661	118092	CRF	C0007947	1617	Haemophilus ducreyi	C0026780	12594	Parotitis	C0007799	6839	Ventricles
C0010583	65189	CTX	C0018481	996	Haemophilus ducreyi	C0030583	2094	Parotitis	C0018827	227871	Ventricles
C0238052	1613	CTX	C0242872	228	Hemlock	C0013058	4388	Pharmaceutical	C0004903	4477	WBS
C0006675	1059398	Ca	C0949851	1302	Hemlock	C0031336	6742	Pharmaceutical	C0175702	10568	WBS
C0006754	56794	Ca	C0626201	4209	Heregulin	C0031705	182933	Phosphorus	C0027708	24437	WT1
C0006823	109981	Ca	C0752253	2589	Heregulin	C0080014	14909	Phosphorus	C0148873	24069	WT1
C0006767	7709	Callus	C0019552	26952	Hip	C0017916	8445	Phosphorylase	C0043041	23585	Wasp
C0376154	703	Callus	C0022122	792	Hip	C0917783	362	Phosphorylase	C0258432	12777	Wasp
C0030163	30914	Cardiac pacemaker	C0021487	3203	IA	C0032064	18717	Plague	C0043395	6033	Yellow Fever
C0037189	9587	Cardiac pacemaker	C0022037	6816	IA	C0032066	1632	Plague	C0301508	2841	Yellow Fever
C0007634	55017	Cell	C0021246	63672	INDO	C0011389	39203	Plaque	C0056208	1886	cRNA
C1136359	9493	Cell	C0021247	18357	INDO	C0333463	23512	Plaque	C1321571	1991	cRNA
C0011343	7360	Cement	C0021069	31909	IP	C0005821	408339	Platelet	C0011485	5280	dC
C1706094	12050	Cement	C0021171	2233	IP	C0032181	85633	Platelet	C0012764	3357	dC
C0008354	39019	Cholera	C0020746	19914	Ice	C0026934	21720	Pleuropneumonia	C0001948	268465	drinking
C0008359	9144	Cholera	C0025611	54867	Ice	C0032241	745	Pleuropneumonia	C0684271	14483	drinking
C0008778	68269	Cilia	C0534519	28409	Ice	C0032533	7620	Polymyalgia Rheumatica	C0018064	1569	eCG
C0015422	1920	Cilia	C0022023	96689	Ion	C0039483	17436	Polymyalgia Rheumatica	C1623258	164449	eCG
C0009237	38237	Coffee	C0022024	8402	Ion	C0032821	226683	Potassium	C0023308	1512	lens
C0085952	11005	Coffee	C0022077	23847	Iris	C0162800	8838	Potassium	C0023317	90021	lens
C0009264	154042	Cold	C1001362	651	Iris	C0016538	28969	Projection	C0023318	7839	lens
C0009443	8249	Cold	C0022341	143137	JP	C0033363	499	Projection	C0019829	82988	lymphogranulomatosis
C0024117	346332	Cold	C0031106	3514	JP	C0002893	1540	RA	C0036202	53699	lymphogranulomatosis
C0009563	9774	Compliance	C0022864	46915	LABOR	C0003873	431693	RA	C0022171	9191	pI
C1321605	55901	Compliance	C0043227	53100	LABOR	C0034625	7321	RA	C0812425	5487	pI
C0001614	395	Cortex	C0006147	139863	Lactation	C0035335	38859	RB	C0032009	5034	posterior pituitary
C0007776	324010	Cortex	C0022925	64698	Lactation	C0035930	5042	RB	C0032017	621	posterior pituitary
C0001613	11909	Cortical	C0023008	105230	Language	C0014772	11901	RBC	C0012931	4088	rDNA
C0007776	324010	Cortical	C0033348	4933	Language	C0014792	353039	RBC	C0012933	55357	rDNA
C0022655	16714	Cortical	C0023078	32290	Laryngeal	C0035236	51084	RSV	C0040395	10463	tomography
C0040441	966	Crack	C0023081	917	Laryngeal	C0086943	1304	RSV	C0040405	818770	tomography
C0085163	5507	Crack	C0752045	579	Lawsonia	C0851346	15849	Radiation	C0042615	17713	veterinary
C0002395	567548	DAT	C1068388	199	Lawsonia	C1522449	146445	Radiation	C0206212	1191	veterinary
C0114838	31063	DAT	C0023281	14290	Leishmaniasis	C0035203	99273	Respiration			
C0025923	3665	DBA	C1548483	1406	Leishmaniasis	C0282636	20581	Respiration			
C1260899	3084	DBA	C0024131	2274	Lupus	C0035298	373380	Retinal			

Table 11: Frequency of every sense: the number of occurences of an ambiguous word in Medline (title and abstract) and PubMed Central annotated with MeSH which corresponds to the CUI

Target	Org Years	Filt Years	Org Size	Filt Size	Excluded	%Included	Target	Org Years	Filt Years	Org Size	Filt Size	Excluded	%Included	Target	Org Years	Filt Years	Org Size	Filt Size	Excluded	%Included
AA	1945-2019	1945-2018	303572	303563	9	100.00	ADA	1946-2018	1955-2018	23516	23483	33	99.86	ADH	1948-2018	1975-2018	32148	31931	217	99.32
ADP	1947-2018	1956-2018	55406	55394	12	99.98	ALS	1946-2018	1948-2018	129646	129641	5	100.00	ANA	1950-2018	1962-2018	18246	18229	17	99.91
Adrenal	1945-2018	1945-2018	171633	171633	0	100.00	Ala	1945-2019	1947-2018	81423	81415	8	99.99	Arteriovenous Anastomoses	1945-2018	1950-2018	1602	1593	9	99.44
Astragalus	1946-2018	1947-2018	8888	8885	3	99.97	B-Cell Leukemia	1946-2018	1986-2018	94020	93807	213	99.77	BAT	1945-2018	1946-2018	131806	131803	3	100.00
BLM	1954-2018	1971-2018	61608	61603	5	99.99	BPD	1956-2018	1980-2018	50606	50604	2	100.00	BR	1945-2019	1946-2018	148563	148549	14	99.99
BSA	1947-2019	1952-2018	73012	72991	21	99.97	BSE	1952-2018	1991-2018	21486	21478	8	99.96	Borrelia	1946-2018	1979-2018	45067	44900	167	99.63
Brucella abortus	1943-2018	1946-2018	6659	6658	1	99.98	CAD	1945-2018	1983-2018	172675	171642	1033	99.40	CAM	1945-2018	1981-2018	39178	39020	158	99.60
CCD	1946-2018	1965-2018	3237	3151	86	97.34	CCl4	1945-2018	1946-2018	22962	22960	2	99.99	CDA	1979-2018	1979-2018	7657	7657	0	100.00
CDR	1958-2018	1973-2018	10565	10536	29	99.73	CH	1945-2019	1946-2018	384806	384718	88	99.98	CI	1945-2019	1949-2018	26664	26651	13	99.95
CIS	1946-2018	1972-2018	17176	16707	469	97.27	CLS	1996-2018	1996-2018	1228	1228	0	100.00	CP	1945-2018	1946-2018	86949	86946	3	100.00
CPDD	1960-2018	1971-2018	276767	276713	54	99.98	CRF	1950-2018	1954-2018	193572	193558	14	99.99	CTX	1951-2018	1960-2018	66806	66802	4	99.99
Ca	1945-2019	1945-2018	1226490	1226173	317	99.97	Callus	1947-2018	1972-2018	8609	8412	197	97.71	Cardiac pacemaker	1949-2018	1954-2018	40512	40501	11	99.97
Cell	1945-2018	1969-2018	65388	64510	878	98.66	Cement	1946-2018	1957-2018	19531	19410	121	99.38	Cholera	1945-2018	1945-2018	48163	48163	0	100.00
Cilia	1945-2018	1950-2018	70204	70189	15	99.98	Coffee	1946-2019	1960-2018	49497	49242	255	99.48	Cold	1945-2019	1945-2018	508626	508623	3	100.00
Compliance	1952-2019	1974-2018	65781	65675	106	99.84	Cortex	1945-2018	1945-2018	324405	324405	0	100.00	Cortical	1945-2018	1945-2018	352633	352633	0	100.00
Crack	1948-2018	1986-2018	6562	6473	89	98.64	DAT	1946-2018	1974-2018	598752	598611	141	99.98	DBA	1947-2018	1972-2018	6809	6749	60	99.12
DDD	1946-2018	1962-2018	3167	3121	46	98.55	DDS	1946-2018	1960-2018	66853	66757	96	99.86	DE	1945-2019	1945-2018	82423	82422	1	100.00
DI	1945-2018	1946-2018	31685	31684	1	100.00	DON	1948-2018	1975-2018	2313	2280	33	98.57	Digestive	1945-2019	1945-2018	36241	36240	1	100.00
EGG	1945-2019	1945-2018	100543	100534	9	99.99	EM	1945-2018	1946-2018	94994	94993	1	100.00	EMS	1945-2018	1967-2018	42361	42194	167	99.61
ERP	1948-2018	1956-2018	84901	84888	13	99.98	ERUPTION	1944-2018	1945-2018	18305	18303	2	99.99	Eels	1946-2018	1951-2018	7588	7583	5	99.93
Epi	1945-2018	1945-2018	116624	116624	0	100.00	Erythrocytes	1945-2019	1945-2018	364943	364940	3	100.00	Exercises	1945-2019	1945-2018	654282	654257	25	100.00
FA	1945-2018	1945-2018	85449	85449	0	100.00	FTC	1952-2018	1982-2018	7443	7432	11	99.85	Familial Adenomatous Polyposis	1947-2018	1986-2018	47392	47303	89	99.81
Fe	1945-2019	1945-2018	572716	572705	11	100.00	Fish	1945-2019	1945-2018	305154	305099	55	99.98	Follicle	1945-2018	1949-2018	71646	71632	14	99.98
Follicles	1945-2018	1949-2018	127371	127354	17	99.99	GAG	1947-2018	1949-2018	50174	50167	7	99.99	Ganglion	1945-2018	1946-2018	17719	17718	1	99.99
Gas	1945-2018	1945-2018	55642	55642	0	100.00	Glycoside	1945-2018	1946-2018	25627	25626	1	100.00	HCl	1945-2018	1946-2018	20316	20315	1	100.00
HGF	1965-2019	1984-2018	432707	432677	30	99.99	HHV 8	1946-2018	1953-2018	96870	96841	29	99.97	HIV	1985-2019	1985-2018	1750664	1750634	30	100.00
HPS	1994-2018	1994-2018	4275	4275	0	100.00	HR	1945-2018	1947-2018	290905	290900	5	100.00	Haemophilus ducreyi	1945-2018	1977-2018	2749	2613	136	95.05
Hemlock	1954-2018	2002-2018	1566	1530	36	97.70	Heregulin	1992-2018	1992-2018	6798	6798	0	100.00	Hip	1945-2018	1946-2018	27747	27744	3	99.99
IA	1945-2018	1946-2018	10020	10019	1	99.99	INDO	1946-2018	1949-2018	82031	82029	2	100.00	IP	1947-2018	1986-2018	34267	34142	125	99.64
Ice	1945-2019	1946-2018	103194	103190	4	100.00	Ion	1945-2018	1945-2018	105091	105091	0	100.00	Iris	1944-2018	1945-2018	24499	24498	1	100.00
JP	1945-2018	1946-2018	146654	146651	3	100.00	LABOR	1945-2018	1945-2018	100015	100015	0	100.00	Lactation	1945-2019	1945-2018	204600	204561	39	99.98
Language	1945-2018	1946-2018	110164	110163	1	100.00	Laryngeal	1945-2018	1945-2018	33207	33207	0	100.00	Lawsonia	1999-2018	2000-2018	781	778	3	99.62
Leishmaniasis	1945-2018	1945-2018	15696	15696	0	100.00	Lupus	1945-2018	1945-2018	257968	257968	0	100.00	MAF	1980-2018	1980-2018	2603	2603	0	100.00
MBP	1972-2018	1973-2018	52691	52688	3	99.99	MCC	1953-2018	1988-2018	13026	13025	1	99.99	MHC	1975-2018	1978-2018	62118	62115	3	100.00
MRS	1947-2018	1950-2018	47126	47125	1	100.00	Malaria	1945-2019	1945-2018	358254	358235	19	99.99	Medullary	1945-2018	1946-2018	33690	33687	3	99.99
Milk	1945-2019	1945-2018	305750	305728	22	99.99	Moles	1945-2018	1946-2018	15138	15135	3	99.98	Murine sarcoma virus	1970-2015	2019-2018	1940	0	1940	0.00
NBS	1945-2019	1947-2018	81867	81864	3	100.00	NEUROFIBROMATOSIS	1945-2018	1990-2018	12180	11716	464	96.19	NM	1946-2018	1946-2018	10239	10239	0	100.00
NPC	1959-2018	1998-2018	21609	21594	15	99.93	Nurse	1945-2018	1945-2018	239675	239675	0	100.00	Nursing	1945-2018	1945-2018	172243	172243	0	100.00
OCD	1946-2018	1975-2018	57668	57464	204	99.65	OH	1945-2019	1946-2018	37953	37951	2	99.99	ORI	1983-2018	1993-2018	15381	15375	6	99.96
Orf	1946-2019	1980-2018	86612	86489	123	99.86	PAC	1949-2018	1995-2018	1360	1353	7	99.49	PAF	1979-2018	1979-2018	55654	55654	0	100.00
PCA	1945-2019	1957-2018	84955	84911	44	99.95	PCB	1970-2019	1971-2018	96634	96610	24	99.98	PCD	1946-2019	1971-2018	1240782	1240583	199	99.98
PCP	1951-2018	1972-2018	28148	28076	72	99.74	PEP	1950-2018	1971-2018	6007	5989	18	99.70	PHA	1949-2018	1975-2018	21563	21411	152	99.30
POL	1946-2019	1946-2018	38355	38354	1	100.00	PR	1945-2018	1945-2018	76865	76865	0	100.00	PVC	1946-2019	1974-2018	18101	17934	167	99.08
Parotitis	1945-2018	1945-2018	14688	14688	0	100.00	Pharmaceutical	1945-2018	1945-2018	11130	11130	0	100.00	Phosphorus	1945-2019	1945-2018	197900	197842	58	99.97
Phosphorylase	1946-2018	1971-2018	9398	8807	591	93.71	Plague	1944-2018	1945-2018	20350	20349	1	100.00	Plaque	1946-2018	1950-2018	62726	62715	11	99.98
Platelet	1945-2019	1945-2018	494007	493972	35	99.99	Pleuropneumonia	1945-2018	1945-2018	22465	22465	0	100.00	Polymyalgia Rheumatica	1946-2018	1957-2018	25082	25056	26	99.90
Potassium	1945-2019	1945-2018	235537	235521	16	99.99	Projection	1946-2018	1970-2018	29708	29468	240	99.19	RA	1945-2018	1945-2018	440554	440554	0	100.00
RB	1946-2018	1947-2018	43903	43901	2	100.00	RBC	1945-2019	1945-2018	364943	364940	3	100.00	RSV	1948-2018	1955-2018	52397	52388	9	99.98
Radiation	1944-2018	1945-2018	162296	162294	2	100.00	Respiration	1945-2018	1945-2018	119854	119854	0	100.00	Retinal	1944-2018	1945-2018	383905	383903	2	100.00
Root	1945-2019	1945-2018	376231	376203	28	99.99	SARS	1953-2018	2002-2018	62411	62409	2	100.00	SARS-associated coronavirus	1953-2018	2002-2018	62411	62409	2	100.00
SCD	1945-2019	1946-2018	81011	80930	81	99.90	SLS	1945-2019	1971-2018	29857	29779	78	99.74	SPR	1981-2019	1981-2018	39679	39670	9	99.98
SS	1945-2018	1948-2018	14413	14410	3	99.98	STEM	1945-2018	1992-2018	27105	26926	179	99.34	Schistosoma mansoni	1945-2018	1971-2018	21689	21036	653	96.99
Sodium	1945-2019	1945-2018	299276	299274	2	100.00	Staph	1945-2018	1945-2018	37349	37349	0	100.00	Sterilization	1945-2018	1945-2018	19072	19072	0	100.00
Strep	1945-2019	1945-2018	36074	36031	43	99.88	Synapsis	1945-2018	1950-2018	130490	130465	25	99.98	TAT	1945-2018	1975-2018	29858	29590	268	99.10
TEM	1950-2019	2002-2018	39293	38618	675	98.28	THYMUS	1945-2018	1946-2018	120292	120289	3	100.00	TLC	1954-2019	1959-2018	25152	25147	5	99.98
TMJ	1946-2018	1946-2018	29019	29019	0	100.00	TMP	1954-2018	1970-2018	15109	15095	14	99.91	TNC	1980-2018	1980-2018	26711	26711	0	100.00
TNT	1945-2018	1981-2018	23761	23623	138	99.42	TPA	1962-2018	1975-2018	142421	142419	2	100.00	TPO	1954-2018	1974-2018	30730	30690	40	99.87
TRF	1956-2018	1968-2018	48110	48090	20	99.96	TSF	1958-2018	1974-2018	193485	193462	23	99.99	TYR	1945-2019	1945-2018	152701	152698	3	100.00
Tax	1950-2019	1975-2018	124646	124603	43	99.97	Tolerance	1945-2018	1946-2018	92750	92747	3	100.00	Torula	1945-2018	1945-2018	23701	23701	0	100.00
US	1945-2019	1945-2018	804649	804623	26	100.00	Ventricles	1945-2018	1945-2018	234710	234710	0	100.00	WBS	1981-2018	1982-2018	15047	15045	2	99.99
WT1	1946-2018	1946-2018	48506	48506	0	100.00	Wasp	1945-2018	1975-2018	36439	36362	77	99.79	Yellow Fever	1945-2018	1945-2018	8874	8874	0	100.00
cRNA	1947-2018	1975-2018	3902	3877	25	99.36	dC	1945-2018	1966-2018	8751	8637	114	98.70	drinking	1946-2018	1946-2018	282948	282948	0	100.00
eCG	1945-2018	1945-2018	166018	166018	0	100.00	lens	1945-2018	1945-2018	99372	99372	0	100.00	lymphogranulomatosis	1945-2018	1945-2018	136687	136687	0	100.00
pI	1945-2018	1975-2018	14767	14678	89	99.40	posterior pituitary	1946-2018	1946-2018	5655	5655	0	100.00	rDNA	1974-2019	1976-2018	59449	59445	4	99.99
sex factor	1946-2018	2019-2018	1604	0	1604	0.00	tomography	1945-2018	1947-2018	829236	829233	3	100.00							

Table 12: a summary about the final set of the data

which is the probabilistic equivalent of matching G_i with several different clusters. Again, this issue is more likely to happen when there is imbalance between the senses: the majority class G_i is likely to obtain the highest $P(G_i|C_j)$ against several clusters, thus becoming the winning sense for any new instance x predicted with a high probability of the corresponding clusters. The table below shows the mapping matrix for the example.

C_j	G_1	G_2	G_3
C_0	**0.45**	0.24	0.30
C_1	0.02	0.26	**0.71**
C_2	**0.74**	0.01	0.24

The mapping matrix is used to transform the cluster probabilities into sense probabilities. For example, given two instances x and y with respective probability vectors $[C_0 = 1, C_1 = 0, C_2 = 0]$ and $[C_0 = 0, C_1 = 0, C_2 = 1]$, multiplying these vectors with the above matrix gives G_1 as the higihest probability in both cases. This means that this evaluation method considers the two instances as both predicted as G_1, even though they belong to different clusters.