

文献紹介ゼミ

林 秀治

紹介する文献

- Uncertainty Detection in Hungarian Texts
- Veronika Vincze
- Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1844-1853,

概要

- 機械学習で文章の不確実性の検出
- ハンガリー語初の不確実性コーパスの公開
- ハンガリー語初のテキストの不確実性検出

背景

- 不確実性の検出が不可欠なタスクがある
 - 医療に関する情報では否定や不確実な情報を区別する必要がある
- テキストから不確実性を示すワードを探すことで検出可能
- 既存の研究はほぼ英語

hUnCertainty Corpus

- 二つのドメインを人手でアノテーション
- ハンガリーWikipediaからランダムに1081段落を抽出(9,722文、180,000トークン)
- Hungarian news portalの事件の記事300件(5,481文、94,000トークン)

不確実性の分類

- Szarvas et al. (2012) と Vincze (2013) に従って不確実性の分類
- Semantic uncertainty: “It may be raining”は本当に雨が降っているか判断できる情報がない
 - Epistemic: 世界知識に基づいて判断できない場合
 - Condition: 仮定・条件文
 - Investigation: 研究報告(～について調べた)など
 - Doxastic: ～と信じている など

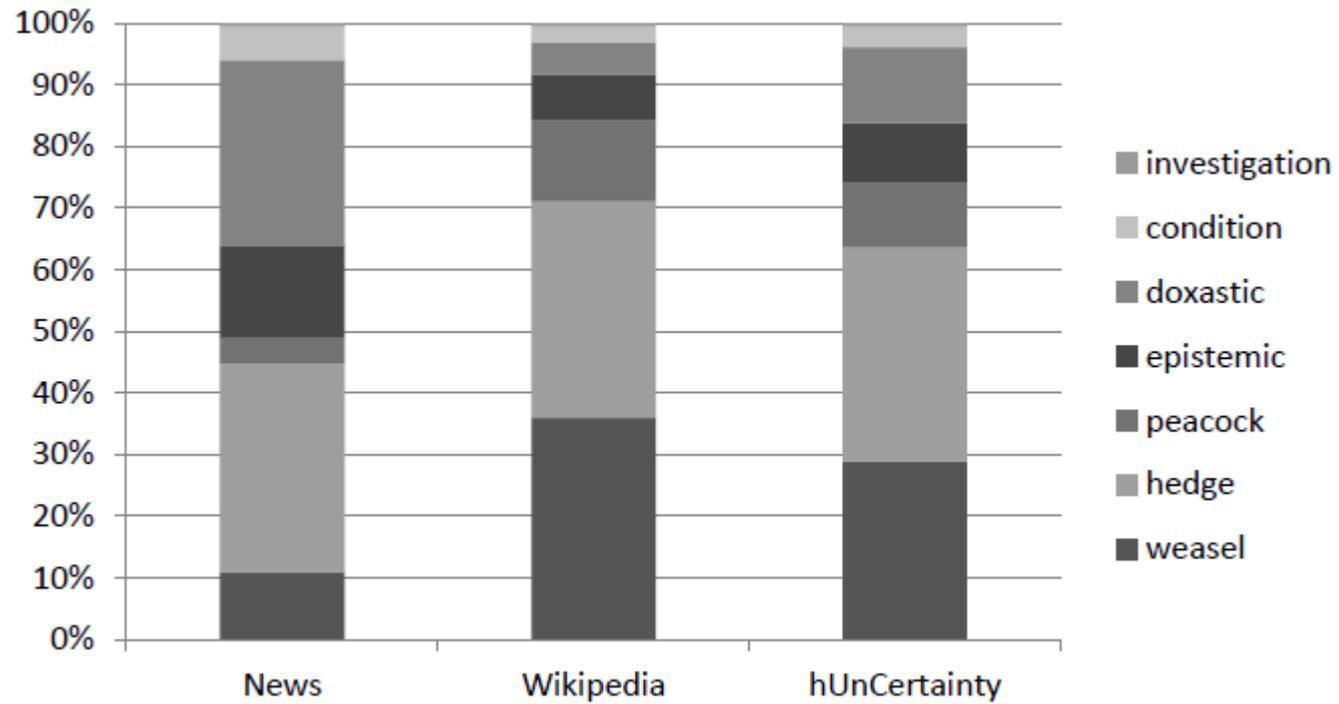
不確実性の分類

- Discourse-level: 会話での不確実性
 - Weasel: ~思うなどの表現
 - Hedge: 一般的な~などの表現
 - Peacock: 誇張表現など

ハンガリー語の不確実性の内訳

Uncertainty cue	Wikipedia		News		Total	
	#	%	#	%	#	%
Weasel	2150	35.95	258	10.93	2408	28.87
Hedge	2100	35.12	800	33.88	2900	34.77
Peacock	788	13.18	94	3.98	882	10.57
Discourse-level total	5038	84.25	1152	48.79	6190	74.21
Epistemic	441	7.37	358	15.16	799	9.58
Doxastic	316	5.28	710	30.07	1026	13.30
Conditional	154	2.58	128	5.42	282	3.38
Investigation	31	0.52	13	0.55	44	0.53
Semantic total	942	15.75	1209	51.21	2151	25.79
Total	5980	100	2361	100	8341	100

コーパス毎の内訳



機械学習

- magyarlanc (Zsibrita et al., 2013).でコーパスを形態素・構文解析
- 英語での先行研究に則り、conditional random feilds(CRF)に基づく手法(Lafferty et al. , 2001)を適用

Feature set

- Orthographic features
 - 語が句読点、アラビア数字、大文字か小文字、子音 bi-、“...”を含むか
- Lexical features
 - 英語の注釈つきコーパスから不確実性の手掛かりを集めてハンガリー語に翻訳したもの

実験結果

- Morphological features
 - 品詞と見出し語
- Syntactic features
 - 従属ラベル
- Semantic/pragmatic features
 - 人手で集めた口語表現

ベースライン

- 単純な辞書引き
- 与えられた語がリストの見出し語に一致したときは、リストに登録されている不確実性のタイプを付与

実験結果

- hUnCertainty Corpusをベースラインと機械学習で分類

Type	Dictionary lookup			Machine learning			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Weasel	18.12	35.92	24.09	52.48	30.73	38.76	+34.37	-5.19	+14.68
Hedge	55.10	32.42	40.82	61.26	48.94	54.41	+6.17	+16.52	+13.59
Peacock	21.66	30.77	25.42	32.61	11.88	17.41	+10.95	-18.89	-8.01
Epistemic	42.46	30.02	35.18	63.18	34.07	44.27	+20.72	+4.04	+9.09
Doxastic	29.30	46.16	35.85	52.42	46.26	49.15	+23.12	+0.10	+13.30
Condition	31.73	62.90	42.18	51.41	25.80	34.35	+19.68	-37.10	-7.83
Micro P/R/F	29.09	35.74	32.07	55.95	37.46	44.87	+26.86	+1.72	+12.80

それぞれのコーパスでの結果

Type	Dictionary lookup			Machine learning			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Weasel	3.24	17.83	5.48	37.50	15.12	21.55	+34.26	-2.71	+16.06
Hedge	53.61	39.05	45.18	61.55	49.69	54.99	+7.94	+10.64	+9.80
Peacock	13.82	31.91	19.29	47.06	8.51	14.41	+33.23	-23.40	-4.88
Epistemic	31.90	20.67	25.08	56.63	39.39	46.46	+24.73	+18.72	+21.37
Doxastic	33.50	37.61	35.43	57.05	51.83	54.32	+23.55	+14.23	+18.88
Condition	35.27	57.03	43.58	54.39	24.22	33.51	+19.12	-32.81	-10.07
Micro P/R/F	23.21	34.17	27.65	57.31	41.93	48.43	+34.10	+7.76	+20.78

Results on the news subcorpus.

Type	Dictionary lookup			Machine learning			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Weasel	26.03	38.50	31.06	59.26	34.74	43.80	+33.23	-3.76	+12.74
Hedge	55.86	29.92	38.97	64.59	50.02	56.38	+8.73	+20.10	+17.41
Peacock	23.29	30.63	26.46	37.85	13.8	20.22	+14.56	-16.83	-6.24
Epistemic	49.57	37.34	42.59	63.95	36.03	46.09	+14.38	-1.31	+3.50
Doxastic	25.24	65.20	36.40	54.31	33.54	41.47	+29.07	-31.66	+5.07
Condition	29.66	67.74	41.26	47.12	31.61	37.84	+17.46	-36.13	-3.42
Micro P/R/F	32.28	36.40	34.21	59.70	37.5	46.06	+27.42	+1.10	+11.85

Results on the Wikipedia subcorpus.

それぞれのコーパスでの結果

Type	Dictionary lookup			Machine learning			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Weasel	3.24	17.83	5.48	37.50	15.12	<u>21.55</u>	+34.26	-2.71	+16.06
Hedge	53.61	39.05	45.18	61.55	49.69	54.99	+7.94	+10.64	+9.80
Peacock	13.82	31.91	19.29	47.06	8.51	14.41	+33.23	-23.40	-4.88
Epistemic	31.90	20.67	25.08	56.63	39.39	46.46	+24.73	+18.72	+21.37
Doxastic	33.50	37.61	35.43	57.05	51.83	54.32	+23.55	+14.23	+18.88
Condition	35.27	57.03	43.58	54.39	24.22	33.51	+19.12	-32.81	-10.07
Micro P/R/F	23.21	34.17	27.65	57.31	41.93	48.43	+34.10	+7.76	+20.78

Results on the news subcorpus.

Type	Dictionary lookup			Machine learning			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Weasel	26.03	38.50	31.06	59.26	34.74	<u>43.80</u>	+33.23	-3.76	+12.74
Hedge	55.86	29.92	38.97	64.59	50.02	56.38	+8.73	+20.10	+17.41
Peacock	23.29	30.63	26.46	37.85	13.8	20.22	+14.56	-16.83	-6.24
Epistemic	49.57	37.34	42.59	63.95	36.03	46.09	+14.38	-1.31	+3.50
Doxastic	25.24	65.20	36.40	54.31	33.54	41.47	+29.07	-31.66	+5.07
Condition	29.66	67.74	41.26	47.12	31.61	37.84	+17.46	-36.13	-3.42
Micro P/R/F	32.28	36.40	34.21	59.70	37.5	46.06	+27.42	+1.10	+11.85

Results on the Wikipedia subcorpus.

それぞれのコーパスでの結果

Type	Dictionary lookup			Machine learning			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Weasel	3.24	17.83	5.48	37.50	15.12	21.55	+34.26	-2.71	+16.06
Hedge	53.61	39.05	45.18	61.55	49.69	54.99	+7.94	+10.64	+9.80
Peacock	13.82	31.91	19.29	47.06	8.51	<u>14.41</u>	+33.23	-23.40	-4.88
Epistemic	31.90	20.67	25.08	56.63	39.39	46.46	+24.73	+18.72	+21.37
Doxastic	33.50	37.61	35.43	57.05	51.83	54.32	+23.55	+14.23	+18.88
Condition	35.27	57.03	43.58	54.39	24.22	33.51	+19.12	-32.81	-10.07
Micro P/R/F	23.21	34.17	27.65	57.31	41.93	48.43	+34.10	+7.76	+20.78

Results on the news subcorpus.

Type	Dictionary lookup			Machine learning			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Weasel	26.03	38.50	31.06	59.26	34.74	43.80	+33.23	-3.76	+12.74
Hedge	55.86	29.92	38.97	64.59	50.02	56.38	+8.73	+20.10	+17.41
Peacock	23.29	30.63	26.46	37.85	13.8	<u>20.22</u>	+14.56	-16.83	-6.24
Epistemic	49.57	37.34	42.59	63.95	36.03	46.09	+14.38	-1.31	+3.50
Doxastic	25.24	65.20	36.40	54.31	33.54	41.47	+29.07	-31.66	+5.07
Condition	29.66	67.74	41.26	47.12	31.61	37.84	+17.46	-36.13	-3.42
Micro P/R/F	32.28	36.40	34.21	59.70	37.5	46.06	+27.42	+1.10	+11.85

Results on the Wikipedia subcorpus.

ドメインによる違いの検証

- wikipediaコーパスで訓練し、newsコーパスを評価
- newsコーパスで訓練し、wikipediaコーパスを評価

ドメインによる違いの検証

Type	Cross validation			Indomain ten fold			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Weasel	17.53	19.77	18.58	37.50	15.12	21.55	+19.97	-4.65	+2.97
Hedge	57.40	39.30	46.66	61.55	49.69	54.99	+4.15	+10.39	+8.33
Peacock	22.81	13.83	17.22	47.06	8.51	14.41	+24.25	-5.32	-2.80
Epistemic	50.00	16.76	25.10	56.63	39.39	46.46	+6.63	+22.63	+21.35
Doxastic	46.63	10.70	17.41	57.05	51.83	54.32	+10.43	+41.13	+36.91
Condition	62.96	26.56	37.36	54.39	24.22	33.51	-8.58	-2.34	-3.85
Micro P/R/F	44.48	23.35	30.62	57.31	41.93	48.43	+12.83	+18.58	+17.81

Cross-domain results: Wikipedia → news.

Type	Cross validation			Indomain ten fold			Difference		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
Weasel	71.26	6.87	12.53	59.26	34.74	43.8	-12.00	+27.87	+31.27
Hedge	63.48	26.33	37.22	64.59	50.02	56.38	+1.11	+23.69	+19.16
Peacock	43.14	5.57	9.87	37.85	13.80	20.22	-5.29	+8.23	+10.35
Epistemic	78.65	30.57	44.03	63.95	36.03	46.09	-14.70	+5.46	+2.06
Doxastic	39.55	33.23	36.12	54.31	33.54	41.47	+14.76	+0.31	+5.35
Condition	47.31	28.39	35.48	47.12	31.61	37.84	-0.19	+3.22	+2.36
Micro P/R/F	59.98	18.00	27.68	59.7	37.5	46.06	-0.28	+19.50	+18.38

Cross-domain results: news → Wikipedia.

文レベルでの検出

- タイプに関係なく文レベルでの不確実性の検出
- ドメインに関係なくほぼ一定の性能

Evaluation setting	Precision	Recall	F-score
hUnCertainty 10 fold	62.20	78.06	69.23
News 10 fold	67.38	78.01	72.30
Wikipedia 10 fold	60.32	80.05	68.80
Wikipedia → news	45.88	74.21	56.70
News → Wikipedia	35.73	84.61	50.24