Inconsistency Detection in Semantic Annotation

Nora Hollenstein
Nathan Schneider
Bonnie Webber

THE UNIVERSITY *of* EDINBURGH

Overview

- Related Work
- Introduction
- Hypothesis
- Data sets
  - Multiword Expressions
  - (Supersense Labels)
- Ranking Methods
  - Discrepancy Ranking
  - Entropy Ranking
- Results
- Conclusion

THE UNIVERSITY *of* EDINBURGH

Related Work

- Syntactic Annotation
  - Inconsistency and error detection in POS Tagging and Treebanks
  - Rule-based approaches (e.g. Ule & Simov (2004))
  - Support Vector Machines (e.g. Nakagawa & Matsumoto (2002))
  - Variation n-gram method (e.g. Dickinson & Meurers (2003))
  - Entropy-based error detection (e.g. Nguyen et al. (2015))

- Semantic Annotation
  - Variation n-gram method (Dickinson & Lee (2008))

Introduction

Annotation inconsistencies

Occurrences of same instances with diverging annotations

Annotation errors

Incorrectly annotated instances

Example:

✗ *in addition to*

✗ *strawberry_banana_milkshake*

Linguistically hard cases[1]

Ambiguities

Example:

I missed you last week.

? *missed = verb.stative OR verb.emotion*

[1] Definition from Klebanov and Beigman (2009)

THE UNIVERSITY of EDINBURGH

Hypothesis

- Detect high **frequency types** which are **most likely to contain inconsistencies** in a corpus with semantic annotations
- Annotations of **multiword expressions** and **supersenses**
- **Ranking methods** compared to a random baseline

Reviewing the highest ranked inconsistency candidates will make the corpus considerably more consistent.

Data sets

MULTIWORD EXPRESSIONS

→ at least two words, which act as a single unit

Inconsistencies examples:

take_care OR *take_care_of*
civil_rights OR *civil_rights_issues*
surprise birthday_party
pumpkin spice latte

SUPERSENSE LABELS

→ coarse-grained semantic classes or word senses

Inconsistency example:

*"Humans live on this **world**, a tiny spot in the milky way."*
? *verb.object* OR *verb.location*

Multiword Expressions

STREUSLE 2.0

- 55’000 tokens
- Web reviews
- Schneider et al. (2014)
- Adjudicated labels, joint annotator consensus
- Strong MWEs, weak MWEs
  - *take_advantage*
  - *highly~recommended*

Wiki50 Corpus

- 100’000 tokens
- 50 Wikipedia articles
- Vincze et al. (2011)
- Five specific types of MWEs
  - *crime_scene* (nom. compound)
  - *high_school* (adj. compound)
  - *spill_the_beans* (idoms)
  - *take_a_break* (light verb const.)
  - *set_up* (verb-part. constructions)

THE UNIVERSITY *of* EDINBURGH

Supersense Labels

STREUSLE 2.0

- Size
- Text types
- Schneider & Smith (2015)
- Supersense tagset for WordNet[1]

Twitter data sets

- 19232 tokens
- tweets
- Johannsen et al. (2014)
- Avoided comprehensive annotation guidelines
- Supersense tagset for WordNet[1]

[1)] 41 labels defined by Ciaramita & Johnson (2003)

THE UNIVERSITY *of* EDINBURGH

Supersense Labels

This store (noun.group) is (verb.stative) proof (noun.cognition) that you can fool (verb.social) people (noun.person) with good advertising (noun.act).

[1] 41 labels defined by Ciaramita & Johnson (2003)

Ranking methods

- Discrepancy ranking
- Entropy ranking

Discrepancy Ranking

1. For each type T, count how many times it is annotated as an MWE in the corpus and how many times it was not annotated:

$$T = (annotated : x, not\text{-}annotated : y)$$

2. For each type T, calculate the following weight W:

$$W = |x - y| \cdot x$$

MWEs – Discrepancy Ranking

	STREUSLE				**Wiki50**			
Rank	MWE	x	y	W	MWE	x	y	W
1	highly recommend	30	3	810	called for	7	1	42
2	thank you	26	2	624	whole body cooling	4	1	12
3	have to	27	16	297	religious classes	3	1	6
4	highly recommended	14	1	182	religious instruction	3	1	6
5	a couple	13	2	143	political crisis	3	1	6
6	work with	12	1	132	brand new	3	1	6
7	a bit	16	10	96	looking for	4	3	4
8	a little	12	4	96	left for	2	4	4
9	worked with	10	1	90	one time	1	5	4
10	at least	10	2	80	went on	3	2	3

Supersense – Discrepancy Ranking

	STREUSLE				**Twitter**			
Rank	word	n	m	W	word	n	m	W
1	place	3	185	10051	day	3	84	2156
2	service	6	200	6400	time	5	64	384
3	staff	2	72	2376	year	2	25	263
4	people	3	84	2072	years	2	18	126
5	car	4	86	1763	night	3	19	101
6	time	3	87	1247	people	2	14	84
7	price	3	61	1179	life	2	13	72
8	experience	3	56	887	work	2	13	46
9	years	2	41	759	today	2	10	26
10	job	3	50	700	show	3	11	24

Entropy Ranking

1. For each type T, calculate its probability p (relative frequency) of being annotated and the probability of not being annotated (*1-p*):

$$p = \frac{x}{x + y}$$

2. Then, calculate the entropy H for each type T:

$$H = -\sum_{i} (p_i) log_2 (p_i)$$

MWEs – Entropy Ranking

	STREUSLE				**Wiki50**			
Rank	MWE	x	y	H	MWE	x	y	H
1	have been	1	50	0.14	called for	7	1	0.54
2	to go	1	32	0.20	one time	1	5	0.65
3	the same	1	32	0.20	whole body cooling	4	1	0.72
4	go to	1	28	0.21	went back	1	4	0.72
5	to see	1	27	0.22	religious classes	3	1	0.81
6	to do	1	24	0.24	religious instruction	3	1	0.81
7	want to	1	23	0.25	political crisis	3	1	0.81
8	go back	1	15	0.34	brand new	3	1	0.81
9	highly recommended	14	1	0.35	fell to	1	3	0.81
10	kind of	1	14	0.35	left for	2	4	0.92

Supersenses – Entropy Ranking

	STREUSLE				Twitter			
Rank	word	n	m	H	word	n	m	H
1	prices	2	36	0.18	people	2	14	0.37
2	area	2	35	0.19	life	2	12	0.39
3	pizza	2	26	0.23	year	2	24	0.40
4	price	3	61	0.24	day	3	84	0.42
5	doctor	2	25	0.24	today	2	10	0.47
6	staff	2	72	0.25	years	2	18	0.50
7	car	4	86	0.27	brithday	2	8	0.54
8	years	2	41	0.28	night	3	19	0.59
9	salon	2	20	0.29	place	2	7	0.59
10	problem	2	20	0.29	followers	2	7	0.92

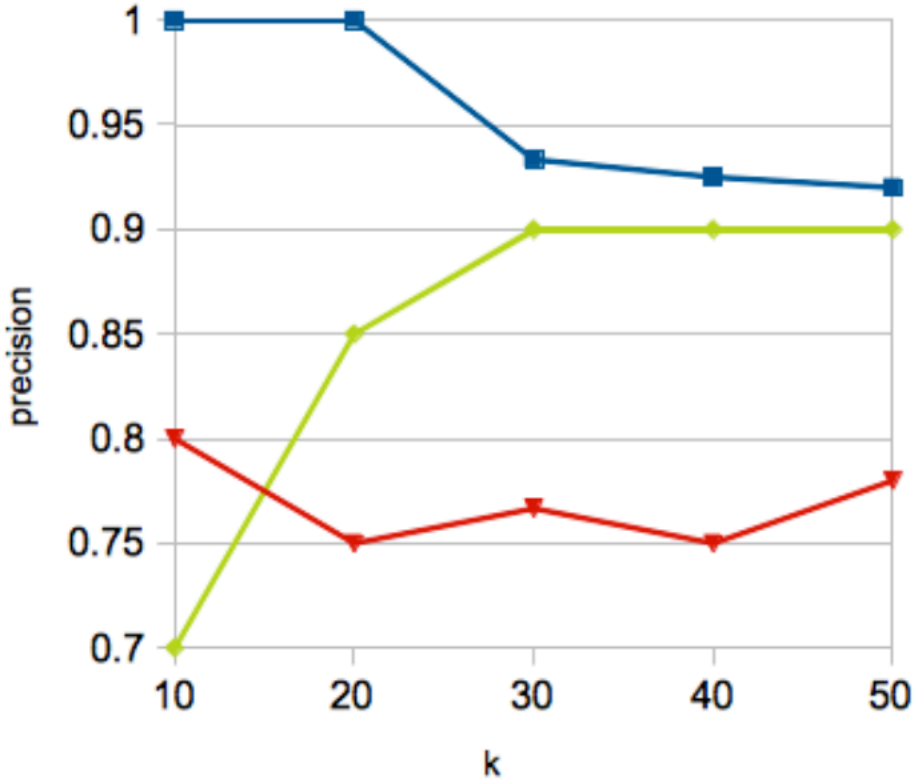
Evaluation

- Manual evaluation (*precision@k*)

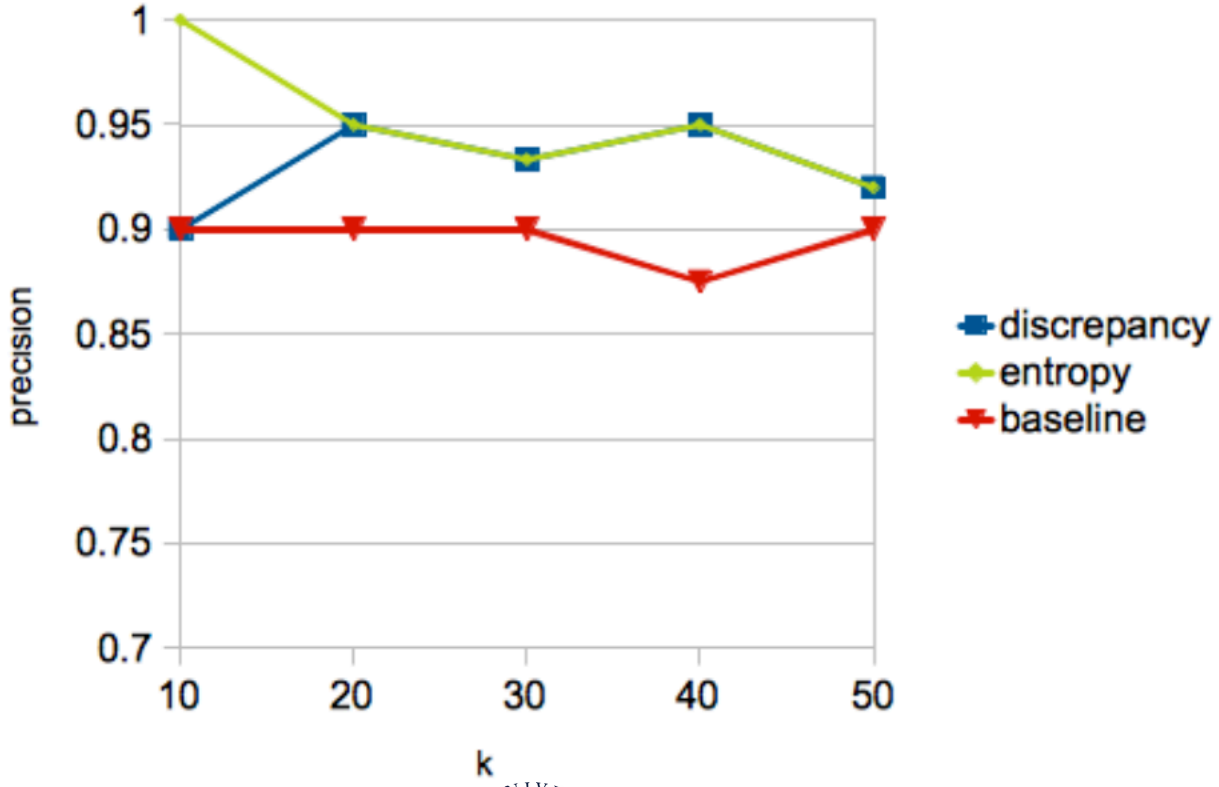
- Significant results over the baseline
  - For both methods
  - For MWEs and supersenses
  - In all four corpora

Results (MWEs)

STREUSLE 2.0



Wiki50 Corpus



THE UNIVERSITY of EDINBURGH

Examples: Inconsistencies

1. → *…the staff up_front will surely **make sure** you get back in time.*
   → *… to **make_sure** everything went well.*

2. → ***Of_course** I couldn't make_it~back in_time.*
   → *Well, unless **of course** the third compressor goes_out.*

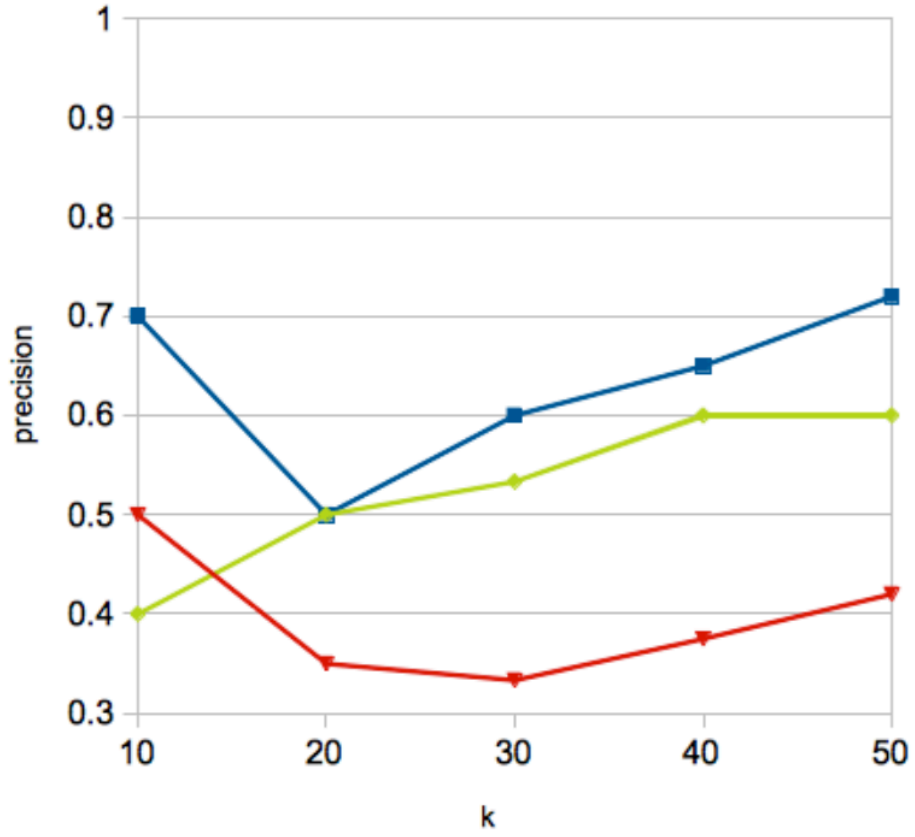
3. →*Thus , he laid ground for a **brand new** way of playing …*
   →*… as well as **brand_new** stages altogether.*

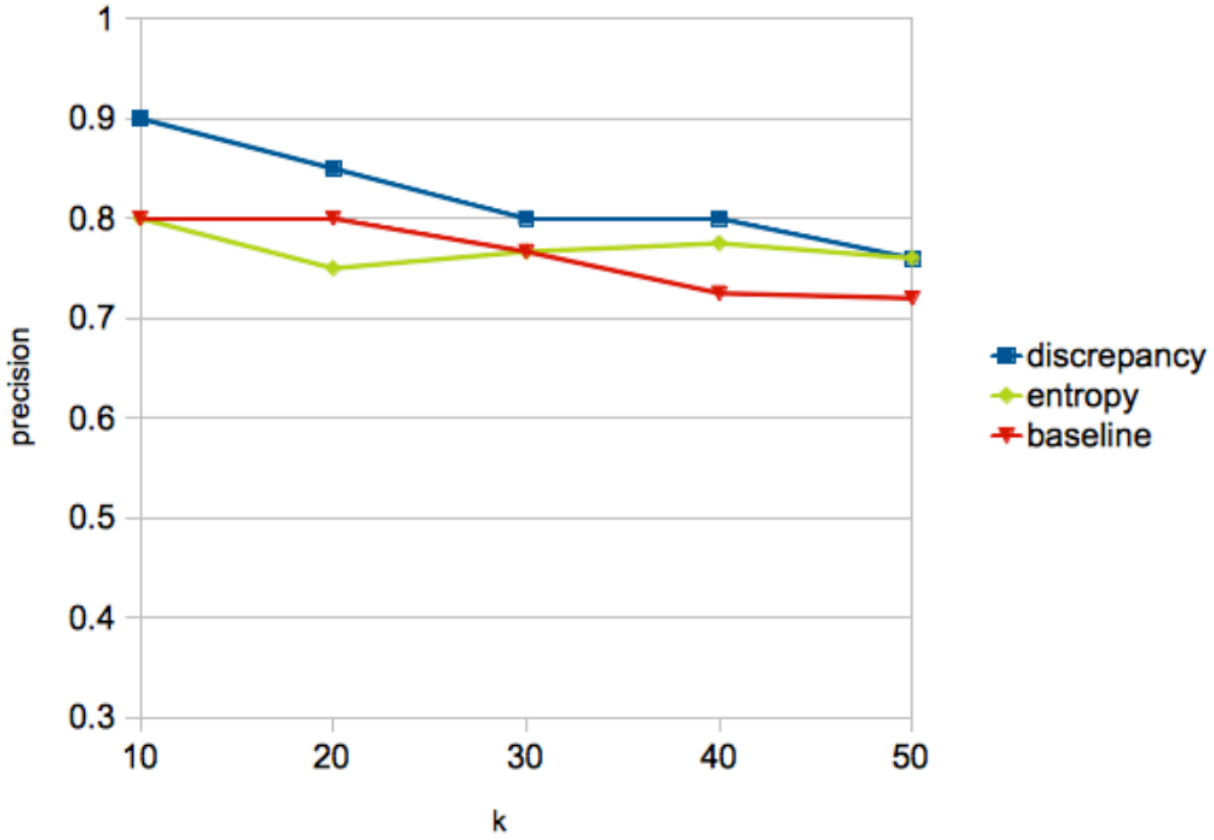
Examples: False Positives

1. → *He has **to go** to school.*
   → *I'll have my coffee **to_go**.*

2. → *I would like to **thank you** for …*
   → ***Thank_you**!*

Results (Supersenses)





Conclusion

- Two new methods for inconsistency detection
  - Applied to multiword expressions and supersense labels
  - Simple methods
  - Easy to apply to other annotation phenomena
- Ranking methods successful in detecting inconsistency candidates
- Future work: integrate these methods into an annotation platform, so that inconsistencies can be caught early

THE UNIVERSITY *of* EDINBURGH

References (1)

B. Beigman Klebanov and E. Beigman. *Difficult cases: From data to learning, and back*, 2009.

N. Schneider, S. Onuffer, N. Kazour, E. Danchik, M. T. Mordowanec, H. Conrad, and N. A. Smith. *Comprehensive annotation of multiword expressions in a social web corpus*. In Proc. of LREC, 2014.

V. Vincze, I. Nagy T., and G. Berend. *Multiword expressions and named entities in the Wiki50 corpus*. In RANLP, pages 289–295, 2011.

N. Schneider and N. A. Smith. *A corpus and model integrating multiword expressions and supersenses*. In Proc. of NAACL-HLT, 2015.

A. Johannsen, D. Hovy, H. M. Alonso, B. Plank, and A. Søgaard. *More or less supervised supersenses tagging of Twitter*. Lexical and Computational Semantics (*SEM 2014), 1, 2014.

References (2)

Dickinson, Markus, and W. Detmar Meurers. *Detecting inconsistencies in treebanks. Proceedings of TLT*. Vol. 3. 2003.

Nguyen, Phuong-Thai, et al. *Vietnamese treebank construction and entropy-based error detection*. *Language Resources and Evaluation* 49.3 (2015): 487-519.

T. Nakagawa and Y. Matsumoto. *Detecting errors in corpora using support vector machines*. In Proceedings of the 19th International Conference on Computational linguistics, volume 1, pages 1–7. Association for Computational Linguistics, 2002.

T. Ule and K. Simov. *Unexpected productions may well be errors*. In LREC, 2004.

M. Ciaramita and M. Johnson. *Supersense tagging of unknown nouns in WordNet*. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pages 168–175. Association for Computational Linguistics, 2003.