Detecting Errors in Semantic Annotation

Markus Dickinson and Chong Min Lee

Indiana University and Georgetown University

LREC 2008, Marrakech, Morocco

Introduction & Motivation

Corpora with semantic annotation are increasingly relevant in natural language processing

- See: Baker et al. (1998); Palmer et al. (2005); Burchardt et al. (2006); Taulé et al. (2005)

Introduction & Motivation

Corpora with semantic annotation are increasingly relevant in natural language processing

- See: Baker et al. (1998); Palmer et al. (2005); Burchardt et al. (2006); Taulé et al. (2005)

Semantic role labeling

- used for tasks such as:
  - information extraction (Surdeanu et al. 2003)
  - machine translation (Komachi et al. 2006)
  - question answering (Narayanan and Harabagiu 2004)

Introduction & Motivation

Corpora with semantic annotation are increasingly relevant in natural language processing

- See: Baker et al. (1998); Palmer et al. (2005); Burchardt et al. (2006); Taulé et al. (2005)

Semantic role labeling

- used for tasks such as:
  - information extraction (Surdeanu et al. 2003)
  - machine translation (Komachi et al. 2006)
  - question answering (Narayanan and Harabagiu 2004)
- requires corpora annotated with predicate-argument structure for training and testing data
  - Gildea and Jurafsky (2002); Xue and Palmer (2004); Toutanova et al. (2005); Pradhan et al. (2005), ...

Introduction & Motivation

Corpora with semantic annotation are increasingly relevant in natural language processing

- See: Baker et al. (1998); Palmer et al. (2005); Burchardt et al. (2006); Taulé et al. (2005)

Semantic role labeling

- used for tasks such as:
  - information extraction (Surdeanu et al. 2003)
  - machine translation (Komachi et al. 2006)
  - question answering (Narayanan and Harabagiu 2004)
- requires corpora annotated with predicate-argument structure for training and testing data
  - Gildea and Jurafsky (2002); Xue and Palmer (2004); Toutanova et al. (2005); Pradhan et al. (2005), ...

Semantically-annotated corpora also have potential as sources of linguistic data for theoretical research

Exploring semantic annotation

Need feedback on annotation schemes:

- difficult to select an underlying theory (see, e.g., Burchardt et al. 2006)
- difficult to determine certain relations, e.g., modifiers (ArgM) in PropBank (Palmer et al. 2005)

Exploring semantic annotation

Need feedback on annotation schemes:

- difficult to select an underlying theory (see, e.g., Burchardt et al. 2006)
- difficult to determine certain relations, e.g., modifiers (ArgM) in PropBank (Palmer et al. 2005)

Need to detect annotation errors, which can:

- harmfully affect training (e.g., van Halteren et al. 2001; Dickinson and Meurers 2005b)
- harmfully affect evaluation (Padro and Marquez 1998; Květŏn and Oliva 2002)

Exploring semantic annotation

Need feedback on annotation schemes:

- difficult to select an underlying theory (see, e.g., Burchardt et al. 2006)
- difficult to determine certain relations, e.g., modifiers (ArgM) in PropBank (Palmer et al. 2005)

Need to detect annotation errors, which can:

- harmfully affect training (e.g., van Halteren et al. 2001; Dickinson and Meurers 2005b)
- harmfully affect evaluation (Padro and Marquez 1998; Květǒn and Oliva 2002)

Little work on automatically detecting errors in semantically-annotated corpora

- Mainly POS and syntactically-annotated corpora (see Dickinson 2005, ch. 1)

Background: the variation n-gram method

Dickinson and Meurers (2003a)

Variation: material occurs multiple times in corpus
with different annotations

Background: the variation n-gram method

Dickinson and Meurers (2003a)

Variation: material occurs multiple times in corpus
with different annotations

Dickinson and Meurers (2003a) introduces the notions

- *variation nucleus*: recurring word with different annotation
- *variation n-gram*: variation nucleus with identical context

and provides an efficient algorithm to compute them.

Background: the variation n-gram method

Dickinson and Meurers (2003a)

Variation: material occurs multiple times in corpus
with different annotations

Dickinson and Meurers (2003a) introduces the notions

- *variation nucleus*: recurring word with different annotation
- *variation n-gram*: variation nucleus with identical context

and provides an efficient algorithm to compute them.

Example: 12-gram with variation nucleus *off*

(1) to ward off a hostile takeover attempt by two European shipping concerns

In the two occurrences of this 12-gram in the WSJ, *off* is

- once annotated as a preposition (IN), and
- once as a particle (RP).

Heuristics for disambigutation

Variation can result from:

- *ambiguity*: different possible labels occur in different corpus occurrences
- *error*: labeling of a string is inconsistent across comparable occurrences

Heuristics for disambigutation

Variation can result from:

- *ambiguity*: different possible labels occur in different corpus occurrences
- *error*: labeling of a string is inconsistent across comparable occurrences

Non-fringe heuristic to detect annotation errors:

- Nuclei found at fringe of *n*-gram more likely to be genuine ambiguities (Dickinson 2005)

Heuristics for disambigutation

Variation can result from:

- *ambiguity*: different possible labels occur in different corpus occurrences
- *error*: labeling of a string is inconsistent across comparable occurrences

Non-fringe heuristic to detect annotation errors:

- Nuclei found at fringe of *n*-gram more likely to be genuine ambiguities (Dickinson 2005)
  - Natural languages favor the use of local dependencies over non-local ones

Error detection for syntactic annotation

Dickinson and Meurers (2003b)

For syntactic annotation, decompose variation nucleus detection into series of runs for all relevant string lengths

Error detection for syntactic annotation

Dickinson and Meurers (2003b)

For syntactic annotation, decompose variation nucleus detection into series of runs for all relevant string lengths

- one-to-one mapping: string $\rightarrow$
  syntactic category label
  (or special label NIL=non-constituent)

Error detection for syntactic annotation

Dickinson and Meurers (2003b)

For syntactic annotation, decompose variation nucleus detection into series of runs for all relevant string lengths

- one-to-one mapping: string →
  syntactic category label
  (or special label NIL=non-constituent)
- perform runs for strings from length 1 to longest constituent in corpus

Error detection for syntactic annotation

Dickinson and Meurers (2003b)

For syntactic annotation, decompose variation nucleus detection into series of runs for all relevant string lengths

- one-to-one mapping: string $\rightarrow$ syntactic category label (or special label NIL=non-constituent)
- perform runs for strings from length 1 to longest constituent in corpus

$\Rightarrow$ High error detection precision for both POS and syntactic annotation

Detecting semantic annotation errors

Method relies on single mapping between text and annotation, but semantic annotation is non-uniform:

(2) [$_{Arg1}$ lending practices] **vary**/vary.01 [$_{Arg2-EXT}$ widely] [$_{ArgM-MNR}$ by location]

Detecting semantic annotation errors

Method relies on single mapping between text and annotation, but semantic annotation is non-uniform:

(2) [$_{Arg1}$ lending practices] **vary**/vary.01 [$_{Arg2-EXT}$ widely] [$_{ArgM-MNR}$ by location]

1. the verb sense

Detecting semantic annotation errors

Method relies on single mapping between text and annotation, but semantic annotation is non-uniform:

(2) [$_{Arg1}$ lending practices] **vary**/vary.01 [$_{Arg2-EXT}$ widely] [$_{ArgM-MNR}$ by location]

1. the verb sense
2. the span of each argument

Detecting semantic annotation errors

Method relies on single mapping between text and annotation, but semantic annotation is non-uniform:

(2) [$_{Arg1}$ lending practices] **vary**/vary.01 [$_{Arg2-EXT}$ widely] [$_{ArgM-MNR}$ by location]

1. the verb sense
2. the span of each argument
3. argument label names

Detecting semantic annotation errors

Method relies on single mapping between text and annotation, but semantic annotation is non-uniform:

(2) [$_{Arg1}$ lending practices] **vary**/vary.01 [$_{Arg2-EXT}$ widely] [$_{ArgM-MNR}$ by location]

1. the verb sense
2. the span of each argument
3. argument label names

Split predicate-argument & verb sense annotation (cf. semantic role labeling, Morante and van den Bosch 2007)

- We focus on argument identification (2) & labeling (3), as these are generally determined by local context

Argument labeling variation

We can view annotation as multiple pairwise relations between a verb & a single argument

Argument labeling variation

We can view annotation as multiple pairwise relations between a verb & a single argument

- While the various arguments are not completely independent, they often have no bearing on each other
  - The manner adverbial *by location* above, for example, does not affect the annotation of *lending practices*

Argument labeling variation

We can view annotation as multiple pairwise relations between a verb & a single argument

- While the various arguments are not completely independent, they often have no bearing on each other
  - The manner adverbial *by location* above, for example, does not affect the annotation of *lending practices*

We define a nucleus as consisting of verb & single argument

- e.g., nuclei for previous sentence: *lending practices vary*, *vary widely*, and *vary by location*

Argument labeling variation

We can view annotation as multiple pairwise relations between a verb & a single argument

- While the various arguments are not completely independent, they often have no bearing on each other
  - The manner adverbial *by location* above, for example, does not affect the annotation of *lending practices*

We define a nucleus as consisting of verb & single argument

- e.g., nuclei for previous sentence: *lending practices vary*, *vary widely*, and *vary by location*
- Semantic annotation involves potentially discontinuous elements (e.g., *vary by location*)
  - use variation *n*-gram algorithm developed for discontinuous syntactic constituency annotation (Dickinson and Meurers 2005a)

Defining a nucleus

Question: What is the label of a nucleus?

Defining a nucleus

Question: What is the label of a nucleus?

- The argument label, e.g., Arg0?
  - Not sufficient: could have the same label, but identify arguments differently

Defining a nucleus

Question: What is the label of a nucleus?

- The argument label, e.g., Arg0?
  - Not sufficient: could have the same label, but identify arguments differently
- Include position of verb in the nucleus
  - e.g., the label of the nucleus *vary widely* is ArgM-MNR-0

Defining a nucleus

Question: What is the label of a nucleus?

- The argument label, e.g., Arg0?
  - Not sufficient: could have the same label, but identify arguments differently
- Include position of verb in the nucleus
  - e.g., the label of the nucleus *vary widely* is ArgM-MNR-0

Can now find errors in argument labeling (e.g., Arg0 vs. Arg1), and in verb identification

Argument identification variation

To find where an argument is unidentified or covers a different stretch of comparable text:

- assign the label NIL to a string not labeled as an argument (cf. Dickinson and Meurers 2005a)

Argument identification variation

To find where an argument is unidentified or covers a different stretch of comparable text:

- assign the label NIL to a string not labeled as an argument (cf. Dickinson and Meurers 2005a)

(3) a. [$_{Arg1}$ net income in its first half] **rose** 59 %

b. [$_{Arg1}$ net income] in its first half **rose** 8.9 %

net income in its first half rose:

- In (3a), assigned label Arg1-6
- In (3b), assigned label NIL

Argument identification variation

To find where an argument is unidentified or covers a different stretch of comparable text:

- assign the label NIL to a string not labeled as an argument (cf. Dickinson and Meurers 2005a)

(3) a. [$_{Arg1}$ net income in its first half] **rose** 59 %

b. [$_{Arg1}$ net income] in its first half **rose** 8.9 %

net income in its first half rose:

- In (3a), assigned label Arg1-6
- In (3b), assigned label NIL

NB: We also recode phrasal verbs as PV relations, to identify variation in phrasal verb identification.

Heuristics for disambiguating strings

Need context to find inconsistent nuclei. Some options:

- Require no identical context of nuclei
  $\rightarrow$ this lack of heuristic gives many false positives

Heuristics for disambiguating strings

Need context to find inconsistent nuclei. Some options:

- Require no identical context of nuclei
  - $\rightarrow$ this lack of heuristic gives many false positives
- Require one word of identical context around every word in nucleus (Dickinson and Meurers 2005a)
  - $\rightarrow$ this “shortest non-fringe” heuristic is very strict

Heuristics for disambiguating strings

Need context to find inconsistent nuclei. Some options:

- Require no identical context of nuclei
  - → this lack of heuristic gives many false positives
- Require one word of identical context around every word in nucleus (Dickinson and Meurers 2005a)
  - → this “shortest non-fringe” heuristic is very strict

We explore another heuristic, in order to increase recall:

- The *argument context heuristic* requires context only around the argument

Heuristics for disambiguating strings

Need context to find inconsistent nuclei. Some options:

- Require no identical context of nuclei
  - → this lack of heuristic gives many false positives
- Require one word of identical context around every word in nucleus (Dickinson and Meurers 2005a)
  - → this "shortest non-fringe" heuristic is very strict

We explore another heuristic, in order to increase recall:

- The *argument context heuristic* requires context only around the argument
- Two main ways that something can be erroneous
  - an error in the labeling (or non-labeling) of the *argument*
  - an error in the identification of the *argument*

Argument context vs. Verb context

- For argument identification, context matters:
  - In (4a), *officials* has no modifier
  - In (4b) *officials* has a modifier

(4) a. Finnair would receive SAS shares valued * at the same amount , [$_{Arg0}$ officials] **said** 0 *T* .

b. ... [$_{Arg0}$ government officials] **said** ...

Argument context vs. Verb context

- For argument identification, context matters:
  - In (4a), *officials* has no modifier
  - In (4b) *officials* has a modifier

(4) a. Finnair would receive SAS shares valued * at the same amount , [$_{Arg0}$ officials] **said** 0 *T* .

b. ... [$_{Arg0}$ government officials] **said** ...

- For verbs, context seems less critical:
  - *substantially reduce* does not depend on what follows

(5) a. That could [$_{Arg2-MNR}$ substantially] **reduce** the value of the television assets .

b. the proposed acquisition could [$_{ArgM-MNR}$ substantially] **reduce** competition ...

Results

We use PropBank as a case study for error detection

Results

We use PropBank as a case study for error detection

Without null element nuclei (cf. Dickinson and Meurers 2003b), we find 43,825 variation nuclei

Results

We use PropBank as a case study for error detection

Without null element nuclei (cf. Dickinson and Meurers 2003b), we find 43,825 variation nuclei

- 369 shortest non-fringe variation nuclei

Results

We use PropBank as a case study for error detection

Without null element nuclei (cf. Dickinson and Meurers 2003b), we find 43,825 variation nuclei

- 369 shortest non-fringe variation nuclei
- 947 variation nuclei with argument context

Results

We use PropBank as a case study for error detection

Without null element nuclei (cf. Dickinson and Meurers 2003b), we find 43,825 variation nuclei

- 369 shortest non-fringe variation nuclei
- 947 variation nuclei with argument context
  - 835 cases involve argument identification variation, i.e., variation with NIL
  - 127 feature variation between labels

Results

We use PropBank as a case study for error detection

Without null element nuclei (cf. Dickinson and Meurers 2003b), we find 43,825 variation nuclei

- 369 shortest non-fringe variation nuclei
- 947 variation nuclei with argument context
  - 835 cases involve argument identification variation, i.e., variation with NIL
  - 127 feature variation between labels

From this set of 947 variations, we sampled 100 cases

- 69% point to inconsistencies, or errors

Results

We use PropBank as a case study for error detection

Without null element nuclei (cf. Dickinson and Meurers 2003b), we find 43,825 variation nuclei

- 369 shortest non-fringe variation nuclei
- 947 variation nuclei with argument context
  - 835 cases involve argument identification variation, i.e., variation with NIL
  - 127 feature variation between labels

From this set of 947 variations, we sampled 100 cases

- 69% point to inconsistencies, or errors

Argument context heuristic successfully increases error detection recall, using only very simple information

POS & Syntactic errors

Overwhelming number of inconsistencies arise from lower-layer annotation errors propagating to PropBank

POS & Syntactic errors

Overwhelming number of inconsistencies arise from lower-layer annotation errors propagating to PropBank

- 42% (29/69) of inconsistencies due to POS errors, as only verbs are annotated in PropBank

(6) a. **coming**/VBG [$_{Arg1}$ months] ,

b. coming/JJ months ,

POS & Syntactic errors

Overwhelming number of inconsistencies arise from lower-layer annotation errors propagating to PropBank

- 42% (29/69) of inconsistencies due to POS errors, as only verbs are annotated in PropBank

(6) a. **coming**/VBG [$_{Arg1}$ months] ,

b. coming/JJ months ,

- 19% (13/69) of inconsistencies due to syntactic errors

(7) a. The following ... are tentatively **scheduled** * [$_{Arg2-for}$ [$_{PP}$ for sale]] this week

b. The following ... are tentatively **scheduled** * [$_{Arg2-for}$ [$_{PP}$ for [$_{NP}$ sale this week]]]

POS & Syntactic errors

Overwhelming number of inconsistencies arise from lower-layer annotation errors propagating to PropBank

- 42% (29/69) of inconsistencies due to POS errors, as only verbs are annotated in PropBank

(6) a. **coming**/VBG [$_{Arg1}$ months] ,

b. coming/JJ months ,

- 19% (13/69) of inconsistencies due to syntactic errors

(7) a. The following ... are tentatively **scheduled** * [$_{Arg2-for}$ [$_{PP}$ for sale]] this week

b. The following ... are tentatively **scheduled** * [$_{Arg2-for}$ [$_{PP}$ for [$_{NP}$ sale this week]]]

- Complements inconsistency detection between syntactic & semantic layers (Babko-Malaya et al. 2006)

Variation in the verb

Also can turn up variation in identifying the verb:

(8) a. the dollar 's [$_{ArgM-MNR}$ continued] **strengthening** reduced world-wide sales growth ...

b. the dollar 's **continued** [$_{Arg1}$ strengthening] reduced world-wide sales growth ...

Variation in the verb

Also can turn up variation in identifying the verb:

(8) a. the dollar 's [$_{ArgM-MNR}$ continued] **strengthening** reduced world-wide sales growth ...

b. the dollar 's **continued** [$_{Arg1}$ strengthening] reduced world-wide sales growth ...

Only example we found, occurring for the same tokens

- Assuming only one element is the head, these cases highlight non-traditional aspects of annotation scheme

Limitations

- Some verbs are ambiguous in whether they take arguments and what type of arguments they take

(9) a. [$_{Arg1}$ Analysts] **had** mixed responses

b. [$_{Arg1}$ Analysts] had **expected** Consolidated to post a slim profit ...

Limitations

- Some verbs are ambiguous in whether they take arguments and what type of arguments they take

(9) a. [$_{Arg1}$ Analysts] **had** mixed responses

b. [$_{Arg1}$ Analysts] had **expected** Consolidated to post a slim profit ...

- Much argument identification ambiguity rooted in difficulties resolving syntactic ambiguity

(10) a. **seeking** [$_{Arg1}$ a buyer] [$_{PP}$ for several months]

b. **seeking** [$_{Arg1}$ a buyer for only its shares]

Limitations

- Some verbs are ambiguous in whether they take arguments and what type of arguments they take

(9) a. [$_{Arg1}$ Analysts] **had** mixed responses

b. [$_{Arg1}$ Analysts] had **expected** Consolidated to post a slim profit ...

- Much argument identification ambiguity rooted in difficulties resolving syntactic ambiguity

(10) a. **seeking** [$_{Arg1}$ a buyer] [$_{PP}$ for several months]

b. **seeking** [$_{Arg1}$ a buyer for only its shares]

- Some argument relations depend upon the sense of the verb, which depends upon other arguments of verb

(11) a. [$_{Arg0}$ he] will **return** Kidder to prominence

b. [$_{Arg1}$ he] will **return** to his old bench

Summary and Outlook

Summary:

- Explored applying the variation *n*-gram error detection method to semantic annotation
  - Defined appropriate units of comparison
  - Relaxed the context definition, using the argument context heuristic
- Found lower layer errors to be primary problem

Summary and Outlook

Summary:

- Explored applying the variation *n*-gram error detection method to semantic annotation
  - Defined appropriate units of comparison
  - Relaxed the context definition, using the argument context heuristic
- Found lower layer errors to be primary problem

Outlook:

- Test on additional corpora with potentially more fine-grained labels, e.g., FrameNet
- Increase recall further (cf. Boyd et al. 2007)
- Explore using only heads of arguments for determining label, to sidestep ambiguous argument identification
  - Such a more general representation potentially more useful for identifying variation in sense annotation

References

Babko-Malaya, Olga, Ann Bies, Ann Taylor, Szuting Yi, Martha Palmer, Mitch Marcus, Seth Kulick and Libin Shen (2006). Issues in Synchronizing the English Treebank and PropBank. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*. Sydney, pp. 70–77.

Baker, Collin F., Charles J. Fillmore and John B. Lowe (1998). The Berkeley FrameNet Project. In *Proceedings of ACL-98*. Montreal, pp. 86–90.

Boyd, Adriane, Markus Dickinson and Detmar Meurers (2007). Increasing the Recall of Corpus Annotation Error Detection. In *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT 2007)*. Bergen, Norway, pp. 19–30.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado and Manfred Pinkal (2006). The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of LREC-06*. Genoa.

Dickinson, Markus (2005). Error detection and correction in annotated corpora. Ph.D. thesis, The Ohio State University.

Dickinson, Markus and W. Detmar Meurers (2003a). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of EACL-03*. Budapest, pp. 107–114.

Dickinson, Markus and W. Detmar Meurers (2003b). Detecting Inconsistencies in Treebanks. In *Proceedings of TLT-03*. Växjö, Sweden, pp. 45–56.

Dickinson, Markus and W. Detmar Meurers (2005a). Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of ACL-05*. pp. 322–329.

Dickinson, Markus and W. Detmar Meurers (2005b). Prune Diseased Branches to Get Healthy Trees! How to Find Erroneous Local Trees in a Treebank and Why It Matters. In *Proceedings of TLT-05*. Barcelona.

Gildea, Daniel and Daniel Jurafsky (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics* 28(4), 245–288.

Komachi, Mamoru, Masaaki Nagata and Yuji Matsumoto (2006). Phrase Reordering for Statisitcal Machine Translation Based on Predicate-Argument Structure. In *Proceedings of the International Workshop on Spoken Language Translation*. Kyoto, Japan, pp. 77–82.

Květŏn, Pavel and Karel Oliva (2002). Achieving an Almost Correct PoS-Tagged Corpus. In Petr Sojka, Ivan Kopeček and Karel Pala (eds.), *Text, Speech and Dialogue (TSD)*. Heidelberg: Springer, no. 2448 in Lecture Notes in Artificial Intelligence (LNAI), pp. 19–26.

Morante, Roser and Antal van den Bosch (2007). Memory-Based Semantic Role Labeling of Catalan and Spanish. In *Proceedings of RANLP-07*. pp. 388–394.

Narayanan, Srini and Sanda Harabagiu (2004). Question Answering based on Semantic Structures. In *International Conference on Computational Linguistics (COLING 2004)*. Geneva, Switzerland.

Padro, Lluis and Lluis Marquez (1998). On the Evaluation and Comparison of Taggers: the Effect of Noise in Testing Corpora. In *COLING/ACL-98*.

Palmer, Martha, Daniel Gildea and Paul Kingsbury (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* 31(1), 71–105.

Pradhan, Sameer, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin and Daniel Jurafsky (2005). Support Vector Learning for Semantic Argument Classification. *Machine Learning* 60(1), 11–39.

Surdeanu, Mihai, Sanda Harabagiu, John Williams and Paul Aarseth (2003). Using Predicate-Argument Structures for Information Extraction. In *Proceedings of ACL-03*.

Taulé, M., J. Aparicio, J. Castellví and M.A. Martí (2005). Mapping syntactic functions into semantic roles. In *Proceedings of TLT-05*. Barcelona.

Toutanova, Kristina, Aria Haghighi and Christopher Manning (2005). Joint Learning Improves Semantic Role Labeling. In *Proceedings of ACL-05*. Ann Arbor, Michigan, pp. 589–596.

van Halteren, Hans, Walter Daelemans and Jakub Zavrel (2001). Improving Accuracy in Word Class Tagging through the Combination of Machine Learning Systems. *Computational Linguistics* 27(2), 199–229.

Xue, Nianwen and Martha Palmer (2004). Calibrating Features for Semantic Role Labeling. In Dekang Lin and Dekai Wu (eds.), *Proceedings of EMNLP 2004*. Barcelona, pp. 88–94.