UTSA Computer Science

Inferring Ontology Fragments From Semantic Role Typing of Lexical Variants

Mitra Bokaei Hosseini[1], Travis D. Breaux[2], **Jianwei Niu**[1]

[1]University of Texas at San Antonio (UTSA)

[2]Carnegie Mellon University

Smart Phone Applications (apps)

Protecting User Privacy

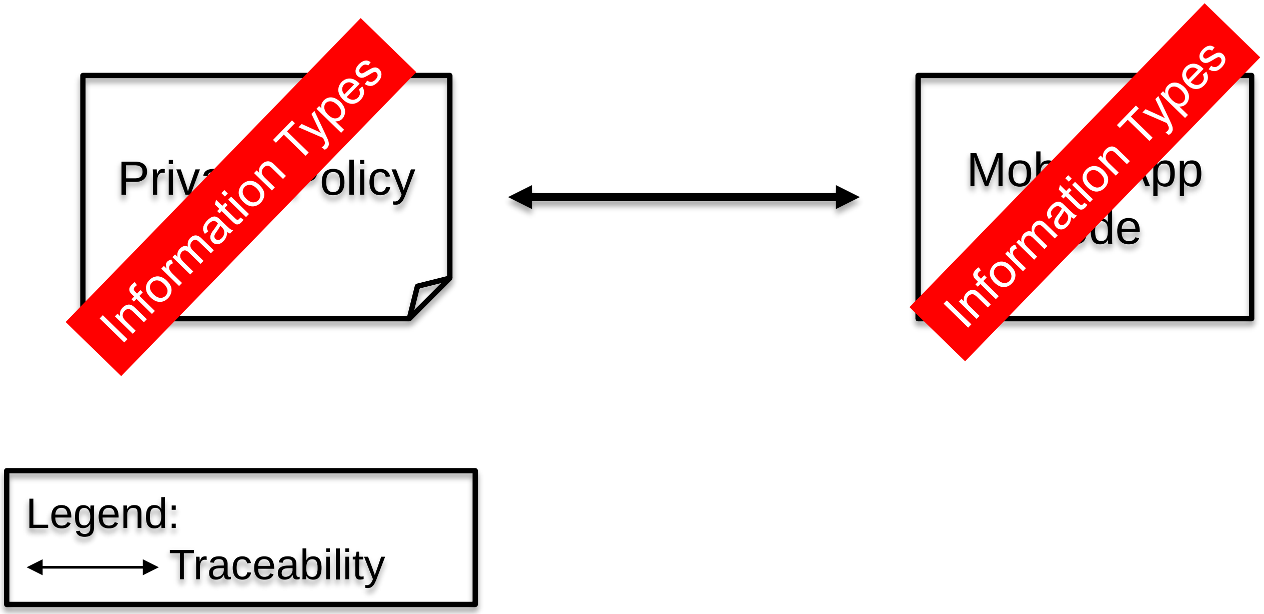
- Growth of access to private information
- Number of apps introduced to the market everyday

To protect users' privacy, we need to identify what information is being collected.

App's Privacy Policy

- Contain critical requirements
- Fulfill legal requirements with respect to General Data Protection Regulation (GDPR) in Europe, and Federal Trade Commission (FTC) Act in the US.
  - California Attorney General's office recommends that policy informs users about what personally identifiable information is collected, used, and shared
- Are expressed in natural language

Trace Links between Policy and Code



Various Interpretation of Data Practices

Adobe Policy Statement: When you activate your Adobe product, we collect certain information about your device, the Adobe product, and your product serial number.

- **Interpretation 1.** ***Mobile device is a kind of device***, then the collection of information also applies to mobile devices (***hypernymy - subsumption***).
- **Interpretation 2.** ***Device has an identifier***, then Adobe may collect device identifier (***meronymy – part-whole***).
- **Interpretation 3.** By use both interpretations (1) and (2), together, we can infer that the collection statement applies to ***mobile device identifier***, using both ***hypernymy*** and ***meronymy***.

Data Collection through Android APIs

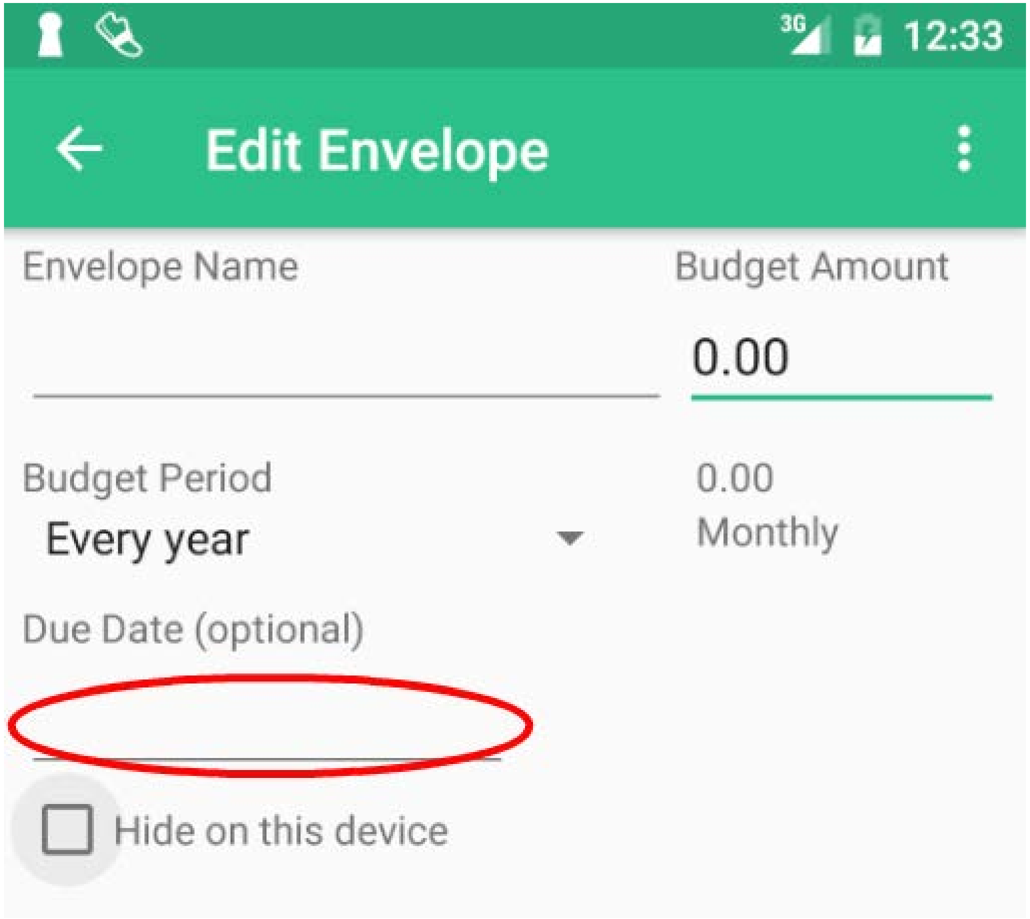
String ANDROID_IDA

64-bit number (as a hex string) that is randomly generated when the user first sets up the device and should remain constant for the lifetime of the user's device. The value may change if a factory reset is performed on the device.

```
import android.provider.Settings.Secure;

private String android_id =
Secure.getString(getContext().getContentResolver(),
Secure.ANDROID_ID);
```

Data Collection through GUI



Research Problems

Abstract and ambiguous information type phrases in privacy policies cause problems in identifying trace links between policy and app code

- Current solutions
  - Manual Ontology Construction
  - Slavin et al. ICSE 2016 and Wang et. al. ICSE 2018
- Proposed solution
  - Developing largely automated techniques and tools to extract semantic relations using syntax

Rocky Slavin, Xiaoyin Wang, Mitra Bikaei Hosseini, James Hester, Ram Krishnan, Jaspret Bhatia, Travis Breaux, and Jianwei Niu, “Toward a framework for detecting privacy policy violations in android application code”. In ICSE 2016

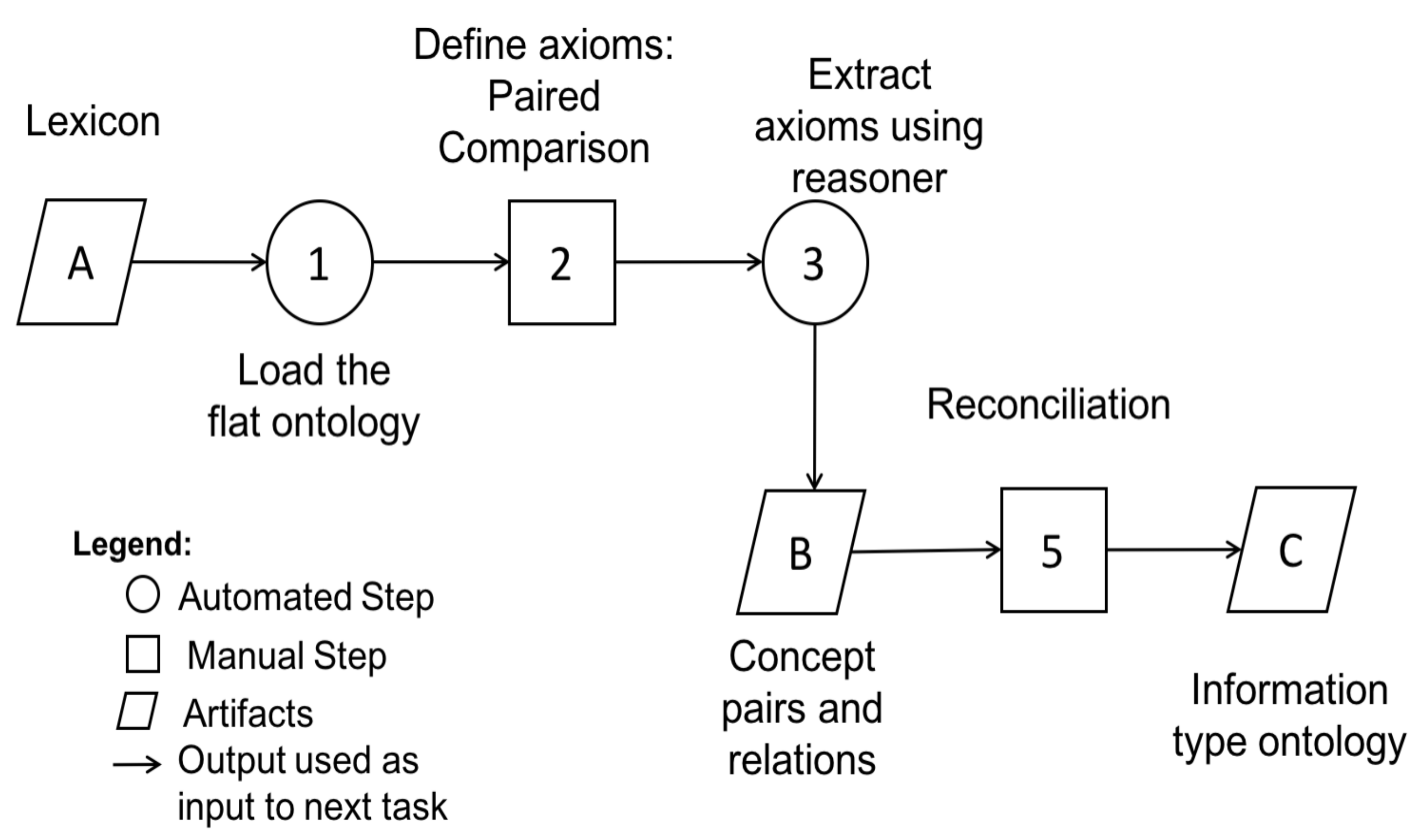
Xiaoyin Wang, Xue Qin, Mitra Bokaei Hosseini, Rocky Slavin, Travis D. Breaux and Jianwei Niu, “GUILeak: Tracing Privacy-Policy Claims on User Input Data for Android Applications”, to appear in ICSE 2018.

Related Work

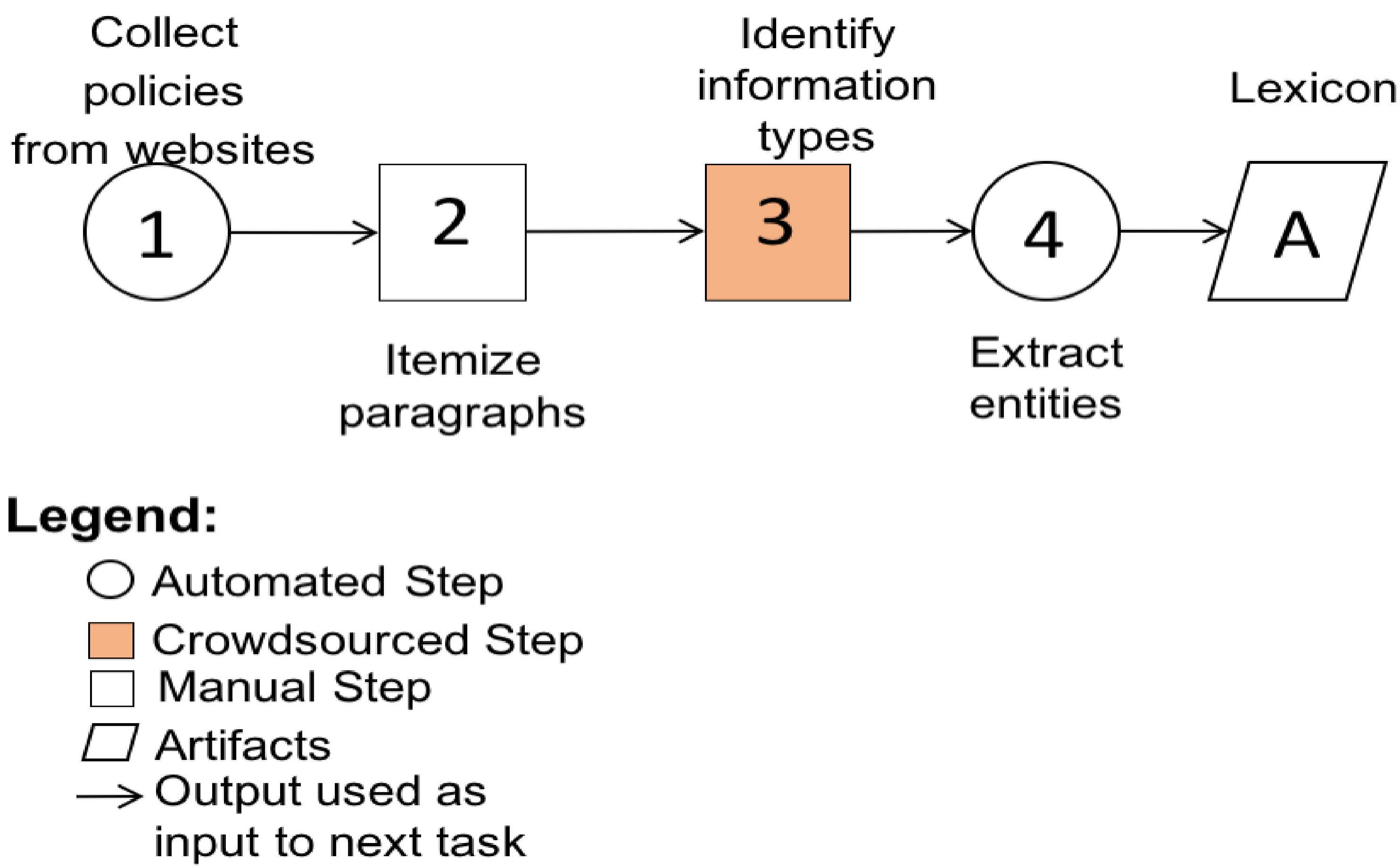
- WordNet: a lexical database on newswire corpus
  - Only contains 14% of 351 information types in our domain
- Existing ontologies: enforcing access control policies, legislative documents, cybersecurity standards
- Our manual ontology construction method use seven Heuristics

Mitra Bokaei Hosseini, Sudarshan Wadkar, Travis D. Breaux, Jianwei Niu, Lexical Similarity of Information Type Hypernyms, Meronyms and Synonyms in Privacy Policies, 2016 AAAI Fall Symposium on Privacy and Language Technologies.

Related Work: Manual Ontology Construction Method



Preparation: Acquiring Privacy Policy Lexicon



Coding Frame for Identifying Information Types

Short Instructions: Select the noun phrases with your mouse cursor and then press one of the following keys to indicate when the noun phrase describes:

- Press 'u' for user provided information - any information that the user explicitly provides to the Tinder or other party
- Press 'a' for automatically collected information - any information that Tinder or another party collects or accesses automatically by the app or website
- Press 'o' for uncertain or unclear - any information that Tinder or another party collects or accesses, and which it is unclear whether the information is provided by the user or by automatic means

In the following paragraph, any pronouns "We" or "Us" refer to Tinder, Inc., and "you" refers to the Tinder user.

Paragraph:

We may collect and store any personal information you provide while using our Service. This may include identifying information, such as your name, address, and email address. We automatically collect information from your browser or device when you visit our Service. This information could include your IP address, device ID and type. We may use information that we collect about you to deliver and improve our products and services, and manage our business.

Manual Ontology Construction: Seven Heuristics for Relation Assignment

- These seven heuristics are the result of grounded analysis of five privacy policies:

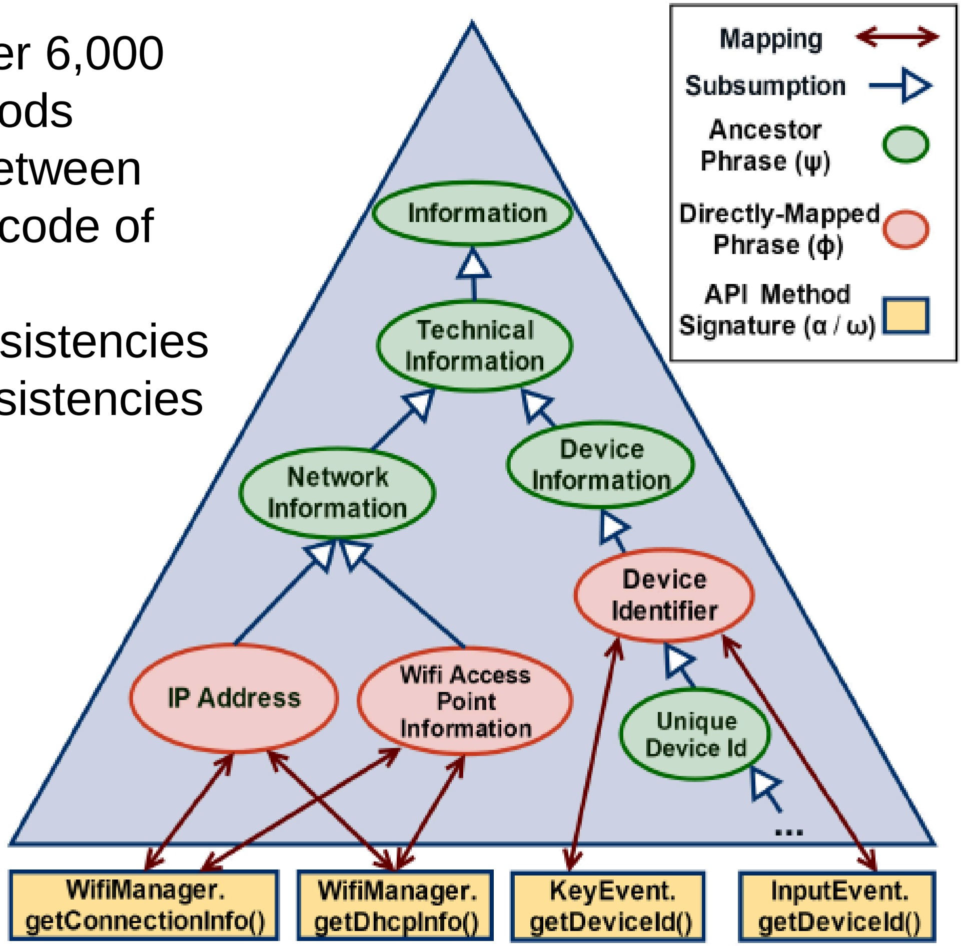
  - Hypernym (H): $C_1 \sqsubseteq C_2$, when concept C_2 is a general category of C_1, e.g., "device" is subsumed by "technology".
  - Meronym (M): $C_1 \sqsubseteq C_2$, when C_1 is a part of C_2, e.g., "internet protocol address" is subsumed by "internet protocol".
  - Attributes (A): $C_1\_C_2 \sqsubseteq C_2$ and $C_1\_C_2 \sqsubseteq C_1\_information$, when the $C_1\_C_2$ phrase contains the C_1 phrase as an attribute or modifier of C_2 phrase, e.g., "unique device identifier" is subsumed by "unique information" and "device identifier".
  - Plural (P): $C_1 \equiv C_2$, when the C_1 phrase is a plural form of the C_2 phrase, e.g., "MAC addresses" is the plural form of "MAC address".
  - Synonym (S): $C_1 \equiv C_2$, when C_1 is a synonym of C_2, e.g., "geo-location" is equivalent to "geographic location".
  - Technology (T): $C_1 \equiv C_1\_information$, when C_1 is a technology, e.g., "device" is equivalent to "device information".
  - Event (E): $C_1 \equiv C_1\_information$, when C_1 is an event, e.g., "usage" is equivalent to "usage information".

Example of Applying Heuristics

LHS Concept	RHS Concept	Heuristic	Analyst1	Analyst2
Device name	Device	Meronymy	SubClass	SubClass
Ads clicked	Usage info	Hypernymy	SubClass	SubClass
Mobile device type	Device type	Modifier	SubClass	None
Tablet	Tablet information	Technology	None	Equivalent
IP address	IP addresses	Plural	Equivalent	Equivalent
Internet protocol address	IP address	Synonym	Equivalent	Equivalent
Usage	Usage Information	Event	Equivalent	Equivalent

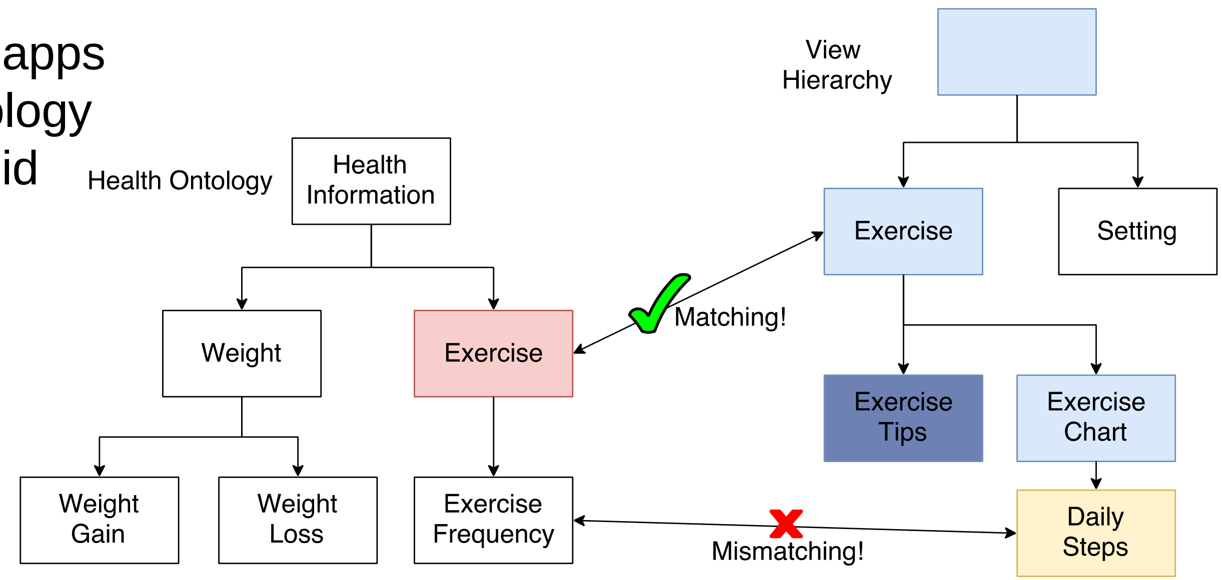
Application of Platform Information Ontology

- Slavin et al. analyzed over 6,000 data producing API methods
- Detect inconsistencies between privacy policies and app code of 477 Android apps
- 344 potential weak inconsistencies
- 58 potential strong inconsistencies



Application of User-provided Information Type Ontologies

- Mapping the View hierarchy of Android apps with the domain ontology
- Analyzing 120 Android apps
- 18 potential weak inconsistencies
- 21 potential strong inconsistencies



Problems with Manual Ontology Construction

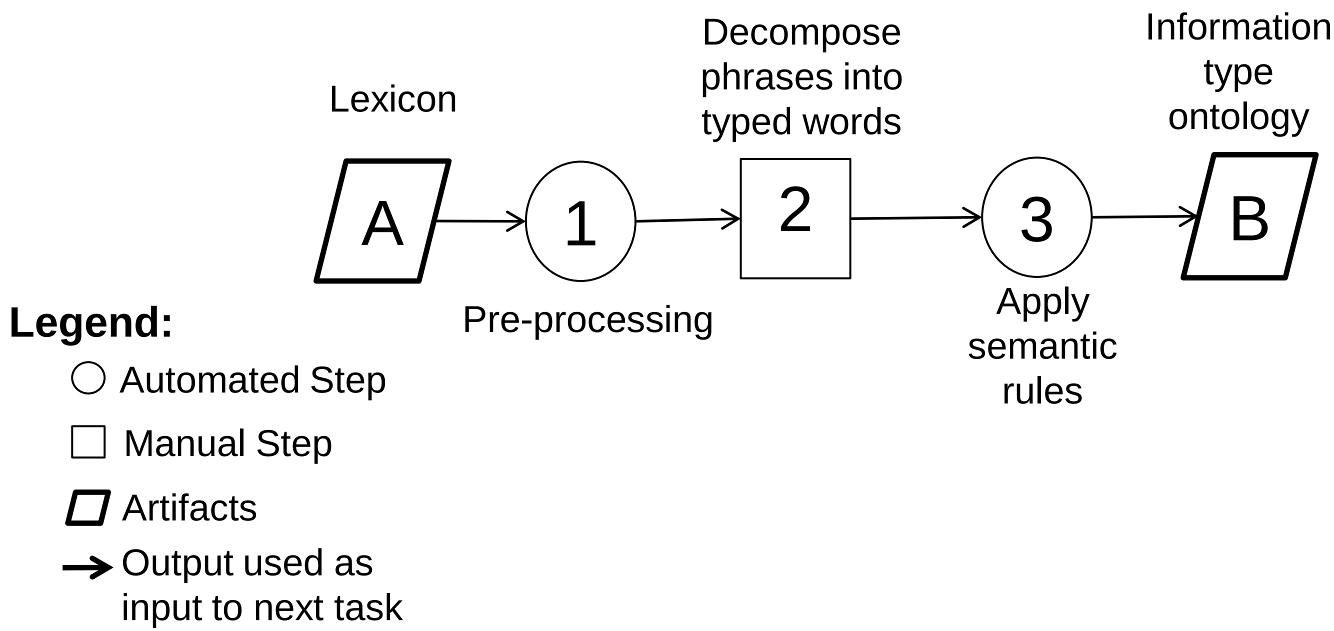
- Requires comparing each information type phrase with every other phrase in the privacy policy lexicon
- Lexicon of 351 phrases results in more than 61,425 comparisons
- Not scalable
- Error prone

Approach: Analyzing Phrases using Syntax

Seven heuristics gave us the following idea

- Analyzing the phrases syntactically
- Example: Mobile device IP address
    - Mobile is modifying device IP address
    - Device IP is the compound noun being modified
    - Address is a property of IP

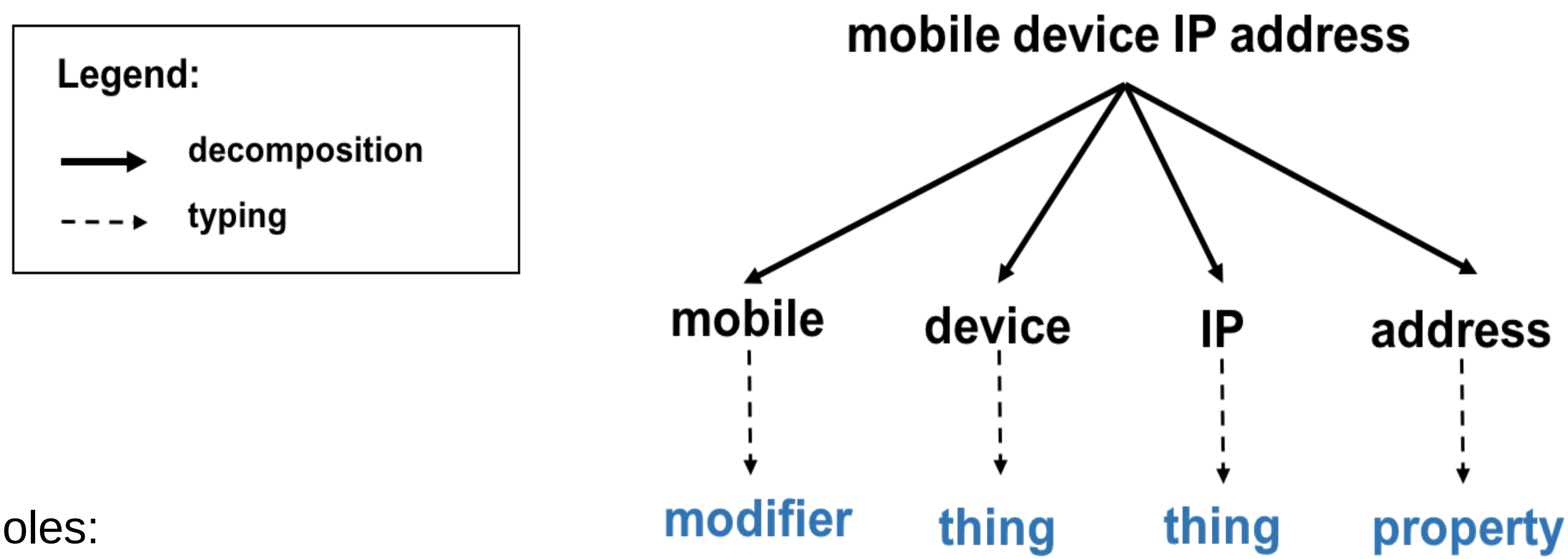
Syntactic Driven Semantic Analysis of Information Types



Lexicon Pre-Processing

- Plural nouns were changed to singular nouns, e.g., “peripherals” is reduced to “peripheral.”
- Possessives were removed, e.g., “device’s information” is reduced to “device information.”
- Suffixes “-related,” “-based,” and “-specific” are removed, e.g., “device-related” is reduced to “device.”
- This reduced the initial lexicon (351 information types) by 16 types to yield a final lexicon with 335 types.

Semantic Role Typing



Roles:
M: Modifier like mobile
E: Event like usage, registration
A: Agent like user
α: Information like information, data
T: Thing like device
P: Property like name, address

Semantic Rules

Applying semantic rules to “mobile device IP address/MTTP”

Role Sequence

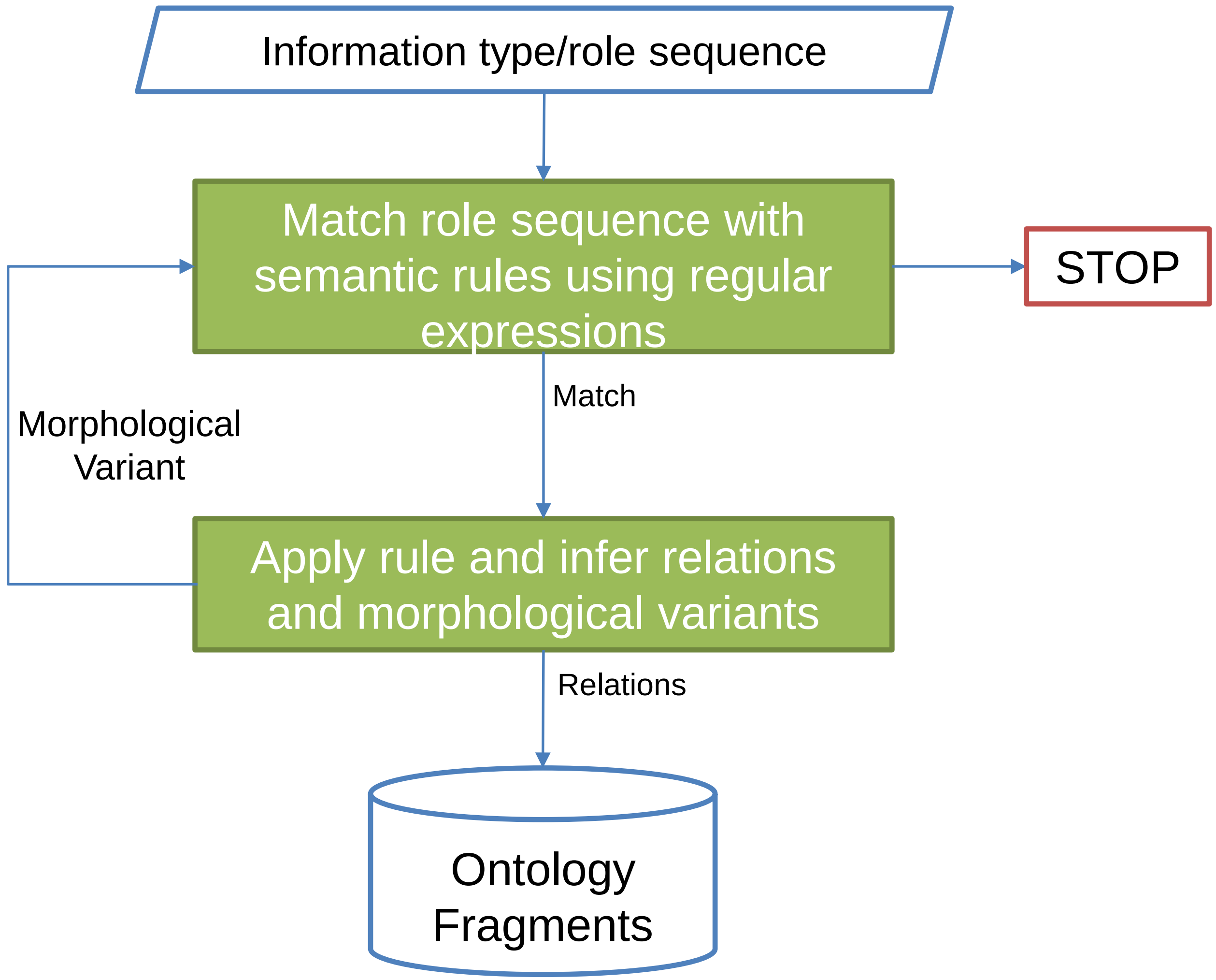
- mobile device IP address is a kind of mobile information
- mobile device IP address is a part of mobile device IP
- device IP address is a part of mobile device IP

Morphological Variants

Morphological Variant

- Given the notion of a lexeme, it is possible to distinguish two kinds of morphological rules.
- Some morphological rules relate to different forms of the same lexeme (inflectional rules). Example: *dog* and *dogs*
- Other rules relate to different lexeme (rules of word formation). Example: compound phrases and words like *dog catcher* or *dishwasher*

Applying Semantic Rules



Evaluation: Experiment Setup

1. **browser** : **web browser type** ☐ click to swap word order
   - ○ is a part of
   - ○ is a kind of
   - ○ is equivalent to
   - ○ is unrelated to
   - ○ unsure or unclear

2. **contact** : **contact list** ☐ click to swap word order
   - ○ is a part of
   - ○ is a kind of
   - ○ is equivalent to
   - ○ is unrelated to
   - ○ unsure or unclear

3. **screen content** : **user content** ☐ click to swap word order
   - ○ is a part of
   - ○ is a kind of
   - ○ is equivalent to
   - ○ is unrelated to
   - ○ unsure or unclear

Survey Details

- 2,365 pairs were surveyed, these pairs all share at least a word.
- We recruited 30 participants to compare each pair using Amazon Mechanical Turk, in which three pairs were shown in one Human Intelligence Task (HIT).
- Qualified participants completed over 5,000 HITs, had an approval rate of at least 97%, and were located in the United States.
- The average time for participants to compare a pair is 11.72 seconds.

Training Stage – Phase 1

- Initial rule set includes 17 semantic rules that are based on the heuristics
- 2,365 information types pairs that share at least a word
- To improve the recall we analyzed false negatives (FNs) and added additional 9 rules to the set

	Semantic Rules
Precision	**0.984**
Recall	**0.221**

Training Stage: Phase 2

- Extended rule set which includes 26 rules
- 2,365 information type pairs that share at least a word
- 477/590 of FNs depend on semantics beyond the scope of the 6-role typology
- 53/590 of FNs were due to individual preference-errors

	Semantic Rules
Precision	**0.996**
Recall	**0.569**

Evaluation: Testing

- Six additional privacy policies resulting in 109 unique information types
- 212 information types pairs that share at least a word
- 44/54 of FNs in the test set depend on semantics beyond the scope of the role typology, example: mobile device and mobile phone
- 7/54 of FNs require introducing new rules

	Semantic Rules
Precision	**1.00**
Recall	**0.593**

Conclusions & Discussions

- The syntax analysis approach is based on the principle of compositionality
  - mobile device IP address
  - mobile device IP address
  - mobile device IP address
- Use syntax to extract semantic relations to facilitate automated ontology construction
- Cannot catch all the type sequences with the rules
- Need to extend our knowledge base to include semantic relations that syntactic analysis cannot infer (relation between phone and device)

Future Work

- Extracting morphological variants using context free grammar and inferring relations using semantic attachments to address the coverage of our current approach
- Using neural networks to expand our knowledge base and infer relations between words/phrases like phone and device