University of Pennsylvania

Weakly-Supervised Acquisition of Labeled Class Instances for Open-Domain Information Extraction

Partha Pratim Talukdar (UPenn) Joseph Reisinger (UT Austin)
Marius Paşca (Google) Deepak Ravichandran (Google)
Rahul Bhagat (USC) Fernando Pereira (Google)

Work done at Google during Summer 2008.

Motivation

- (Class, Instance) pairs (e.g. *(pain killer, aspirin)*) can be useful in many applications e.g. web search.

Motivation

- (Class, Instance) pairs (e.g. *(pain killer, aspirin)*) can be useful in many applications e.g. web search.

- Given an entity/instance, it is often desirable to know its type.

Motivation

- (Class, Instance) pairs (e.g. *(pain killer, aspirin)*) can be useful in many applications e.g. web search.
- Given an entity/instance, it is often desirable to know its type.
- A limited number of classes are not enough:

Motivation

- (Class, Instance) pairs (e.g. *(pain killer, aspirin)*) can be useful in many applications e.g. web search.
- Given an entity/instance, it is often desirable to know its type.
- A limited number of classes are not enough:
  - Web search queries include *active volcanoes* like *Kilauea*, *zoonotic diseases* like *monkeypox* etc., demonstrating general user interest in them.

Motivation

- (Class, Instance) pairs (e.g. *(pain killer, aspirin)*) can be useful in many applications e.g. web search.
- Given an entity/instance, it is often desirable to know its type.
- A limited number of classes are not enough:
  - Web search queries include *active volcanoes* like *Kilauea*, *zoonotic diseases* like *monkeypox* etc., demonstrating general user interest in them.
  - Covering one class at a time (as in standard Named Entity Extraction) is resource intensive and not sufficient.

Motivation

- (Class, Instance) pairs (e.g. *(pain killer, aspirin)*) can be useful in many applications e.g. web search.
- Given an entity/instance, it is often desirable to know its type.
- A limited number of classes are not enough:
  - Web search queries include *active volcanoes* like *Kilauea*, *zoonotic diseases* like *monkeypox* etc., demonstrating general user interest in them.
  - Covering one class at a time (as in standard Named Entity Extraction) is resource intensive and not sufficient.
  - Need open domain extraction involving large number of classes and large number of instances.

Previous Work

Previous Work

- Named Entity Extraction: small number of classes, extensive supervision.

Previous Work

- Named Entity Extraction: small number of classes, extensive supervision.
- (Van Durme and Pasca, AAAI 08): open domain extraction, high precision, low recall: precision drops fast with increasing recall.

Previous Work

- Named Entity Extraction: small number of classes, extensive supervision.
- (Van Durme and Pasca, AAAI 08): open domain extraction, high precision, low recall: precision drops fast with increasing recall.
- Our starting point: extractions from (Van Durme and Pasca, 2008).

Class	Size	Examples of Instances
Book Publishers	70	Crown Publishing, Kluwer Academic, Prentice Hall, Puffin, ...

Objectives

Starting with such automatically extracted (class, instance) pairs:

Objectives

Starting with such automatically extracted (class, instance) pairs:

- Extract additional **instances** for existing **classes**.

Objectives

Starting with such automatically extracted (class, instance) pairs:

- Extract additional **instances** for existing **classes**.
- Identify additional **class labels** for existing **instances**.

Objectives

Starting with such automatically extracted (class, instance) pairs:

- Extract additional **instances** for existing **classes**.
- Identify additional **class labels** for existing **instances**.
- Handle initial pairs from diverse sources and methods.

UNIVERSITY of PENNSYLVANIA

Objectives

Starting with such automatically extracted (class, instance) pairs:

- Extract additional **instances** for existing **classes**.
- Identify additional **class labels** for existing **instances**.
- Handle initial pairs from diverse sources and methods.
- Require minimal human supervision.

Objectives

Starting with such automatically extracted (class, instance) pairs:

- Extract additional **instances** for existing **classes**.
- Identify additional **class labels** for existing **instances**.
- Handle initial pairs from diverse sources and methods.
- Require minimal human supervision.
- Do all these in a scalable manner.

Objectives

Starting with such automatically extracted (class, instance) pairs:

- Extract additional **instances** for existing **classes**.
- Identify additional **class labels** for existing **instances**.
- Handle initial pairs from diverse sources and methods.
- Require minimal human supervision.
- Do all these in a scalable manner.
- Increase coverage (recall) at comparable quality (precision)!

Where do we get instances from?

Where do we get instances from?

- A8: Extractions from unstructured text by (Van Durme and Pasca, AAAI 08).

Where do we get instances from?

- A8: Extractions from unstructured text by (Van Durme and Pasca, AAAI 08).
- WebTables (Cafarella et al., VLDB 2008)

Where do we get instances from?

- A8: Extractions from unstructured text by (Van Durme and Pasca, AAAI 08).
- WebTables (Cafarella et al., VLDB 2008)
  - 154M HTML tables extracted from the web.

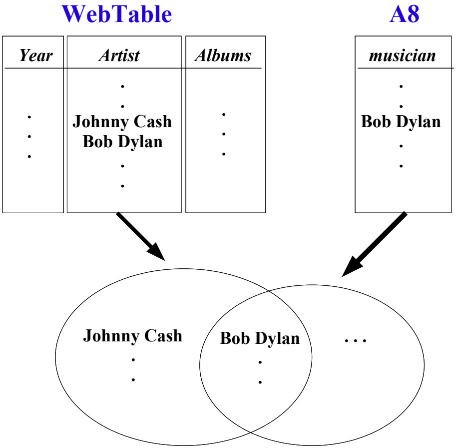
Where do we get instances from?

- A8: Extractions from unstructured text by (Van Durme and Pasca, AAAI 08).
- WebTables (Cafarella et al., VLDB 2008)
  - 154M HTML tables extracted from the web.
  - Rich source of instances, already segmented by webpage creators.

Where do we get instances from?

- A8: Extractions from unstructured text by (Van Durme and Pasca, AAAI 08).
- WebTables (Cafarella et al., VLDB 2008)
  - 154M HTML tables extracted from the web.
  - Rich source of instances, already segmented by webpage creators.
  - Structured text.

Assigning class labels to WebTable instances

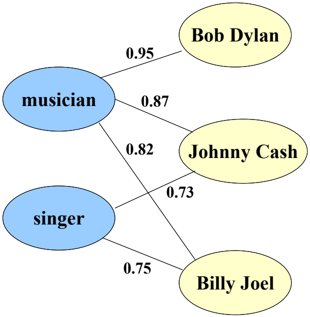


Score (musician, Johnny Cash) = 0.87

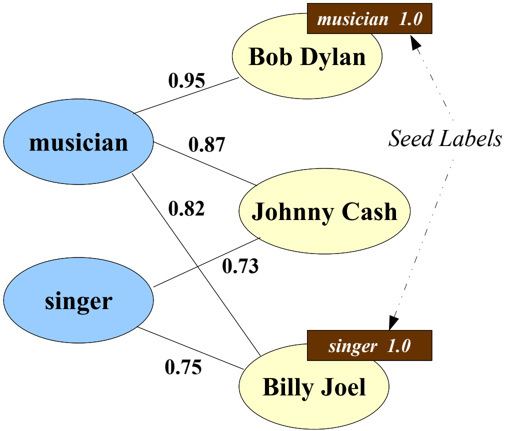
Putting together tuples from first phase extractors

Putting together tuples from first phase extractors

- A graph based representation is used: each tuple from A8 and WebTable is a weighted edge, with nodes representing classes and instances.



Initialization: Seed Labels Marked

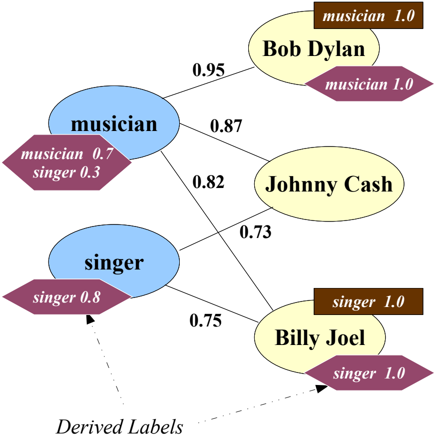


UNIVERSITY of PENNSYLVANIA

Label Propagation: Adsorption (Baluja et al., 2008)

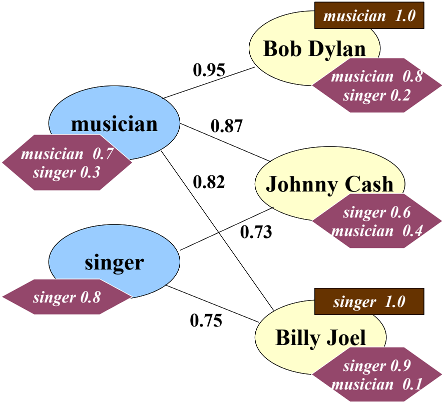
Label Propagation: Adsorption (Baluja et al., 2008)

- After 1 iteration:



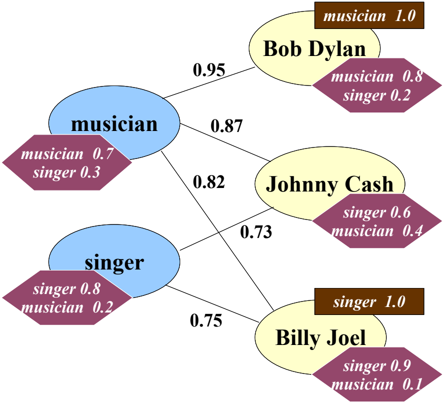
Label Propagation: Adsorption (Baluja et al., 2008)

- After 2 iterations:



Label Propagation: Adsorption (Baluja et al., 2008)

- After 3 iterations:



Experimental Setup

- Dataset **A8**:
    - 924K (class, instance) pairs extracted from 100M web docs.
    - Extracted from *unstructured* text.
    - High precision, low recall.

Experimental Setup

- Dataset **A8**:
  - 924K (class, instance) pairs extracted from 100M web docs.
  - Extracted from *unstructured* text.
  - High precision, low recall.
- Dataset **WT**:
  - 74M unique additional pairs extracted from WebTables.
  - Source of new instances, extracted from *structured* text.
  - Low precision, high recall.

Experimental Setup

- Dataset **A8**:
  - 924K (class, instance) pairs extracted from 100M web docs.
  - Extracted from *unstructured* text.
  - High precision, low recall.
- Dataset **WT**:
  - 74M unique additional pairs extracted from WebTables.
  - Source of new instances, extracted from *structured* text.
  - Low precision, high recall.
- Set of class labels in WT is the same as in A8.

Experimental Setup

- Dataset **A8**:
  - 924K (class, instance) pairs extracted from 100M web docs.
  - Extracted from *unstructured* text.
  - High precision, low recall.
- Dataset **WT**:
  - 74M unique additional pairs extracted from WebTables.
  - Source of new instances, extracted from *structured* text.
  - Low precision, high recall.
- Set of class labels in WT is the same as in A8.
- Graph constructed using A8 + WT had 1.4M nodes and 75M edges. This graph is used in all subsequent experiments.

UNIVERSITY of PENNSYLVANIA

EXPERIMENTS

Experiments

- EXPT 1: Can we find new instances for fixed classes?

Experiments

- EXPT 1: Can we find new instances for fixed classes?
- EXPT 2: For a fixed set of instances, can we assign better class labels?

EXPT 1: Seed (Class, Instance) Pairs

Seed Class	Seed Instances
Book Publishers	Millbrook Press, Academic Press, Springer Verlag, Chronicle Books, Shambhala Publications
NFL Players	Ike Hilliard, Isaac Bruce, Torry Holt, Jon Kitna, Jamal Lewis
Scientific Journals	American Journal of Roentgenology, PNAS, Journal of Bacteriology, American Economic Review, IBM Systems Journal

Table: Classes and seeds used to initialize Adsorption.

UNIVERSITY of PENNSYLVANIA

EXPT 1: Finding new instances for fixed classes

Class	Precision at 100 (non-A8 extractions)
Book Publishers	87.36
Federal Agencies	29.89
NFL Players	94.95
Scientific Journals	90.82
Mammal Species	84.27

Table: Precision of top 100 Adsorption extractions **not** present in A8.

EXPT 1: Finding new instances for fixed classes

Class	Precision at 100 (non-A8 extractions)
Book Publishers	87.36
Federal Agencies	29.89
NFL Players	94.95
Scientific Journals	90.82
Mammal Species	84.27

Table: Precision of top 100 Adsorption extractions **not** present in A8.

Coverage increased at precision level comparable to A8.

New extractions found by Adsorption

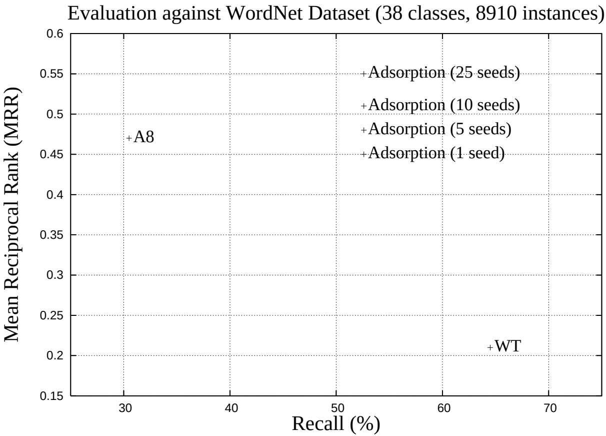
Seed Class	Top Ranked Instances Discovered by Adsorption
Scientific Journals	Journal of Physics, Nature, Structural and Molecular Biology, Sciences Sociales et santé, Kidney and Blood Pressure Research, American Journal of Physiology–Cell Physiology
NFL Players	Tony Gonzales, Thabiti Davis, Taylor Stubblefield, Ron Dixon, Rodney Hannah
Book Publishers	Small Night Shade Books, House of Anansi Press, Highwater Books, Distributed Art Publishers, Copper Canyon Press

Semantically similar class labels found by Adsorption: A Byproduct

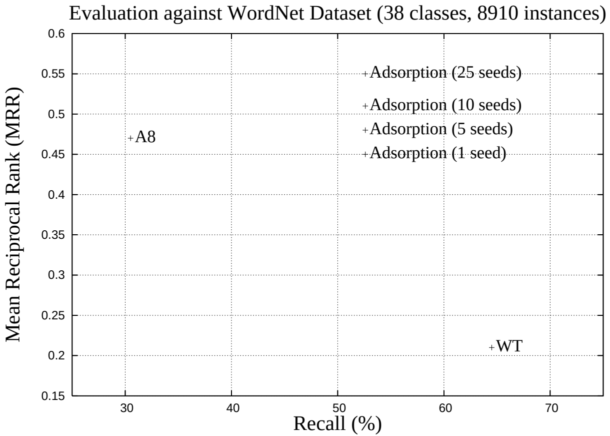
Seed Class	Non-Seed Class Labels Discovered
Book Publishers	small presses, journal publishers, educational publishers, academic publishers, commercial publishers
NFL Players	sports figures, football greats, football players, backs, quarterbacks
Scientific Journals	prestigious journals, peer-reviewed journals, refereed journals, scholarly journals, academic journals

Table: Top class labels ranked by their similarity to a given seed class in Adsorption.

EXPT 2: Class assignment for fixed instances



EXPT 2: Class assignment for fixed instances



Adsorption is able to assign ***better*** class labels to ***more*** instances.

Conclusion

- Demonstrated a scalable graph-based label propagation algorithm.

Conclusion

- Demonstrated a scalable graph-based label propagation algorithm.
- Improved coverage while maintaining adequate precision.

Conclusion

- Demonstrated a scalable graph-based label propagation algorithm.
- Improved coverage while maintaining adequate precision.
- Combined information from two different sources: unstructured and structured texts.

Conclusion

- Demonstrated a scalable graph-based label propagation algorithm.
- Improved coverage while maintaining adequate precision.
- Combined information from two different sources: unstructured and structured texts.
- Future Work:

Conclusion

- Demonstrated a scalable graph-based label propagation algorithm.
- Improved coverage while maintaining adequate precision.
- Combined information from two different sources: unstructured and structured texts.
- Future Work:
  - Class label assignment in context.

Conclusion

- Demonstrated a scalable graph-based label propagation algorithm.
- Improved coverage while maintaining adequate precision.
- Combined information from two different sources: unstructured and structured texts.
- Future Work:
  - Class label assignment in context.
  - Scaling up further!

University of Pennsylvania

Thank You!