Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Alexander Panchenko, Dmitry Ustalov, Stefano Faralli, Simone Paolo Ponzetto, and Chris Biemann

IMPROVING HYPERNYMY EXTRACTION WITH DISTRIBUTIONAL SEMANTIC CLASSES

Introduction

Hypernyms

Examples of hypernymy relations

- **apple** –isa→ **fruit**
- **mangosteen** –isa→ **fruit**

Hypernyms

Examples of hypernymy relations

- **apple** –isa→ **fruit**
- **mangosteen** –isa→ **fruit**

Examples of applications of hypernyms

- question answering [Zhou et al., 2013]
- query expansion [Gong et al., 2005]
- semantic role labelling [Shi & Mihalcea, 2005]

Automatic extraction of hypernyms

Induction of semantic classes

Main contributions

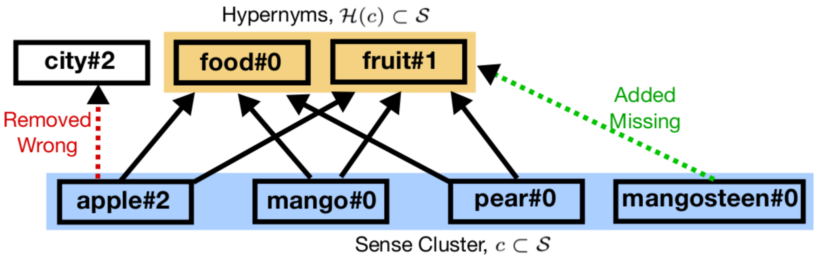
- We show how distributionally-induced **semantic classes** can be helpful for **extracting hypernyms**:

Main contributions

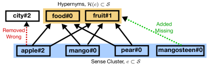
- We show how distributionally-induced **semantic classes** can be helpful for **extracting hypernyms**:
  1. A method for inducing sense-aware semantic classes using distributional semantics;
  2. A method for using the induced semantic classes for filtering noisy hypernymy relations.

Method

Labeled semantic classes



- Post-processing of hypernymy relations using distributionally induced semantic classes;
- A semantic class is a clusters of induced word senses labeled with hypernyms.



- Post-processing of hypernymy relations using distributionally induced semantic classes;
- A semantic class is a clusters of induced word senses labeled with hypernyms.

The word postfix, such as #1, is an ID of an induced sense. The wrong hypernyms outside the cluster labels are removed, while the missing ones not present in the noisy database of hypernyms are added.

Outline of our approach

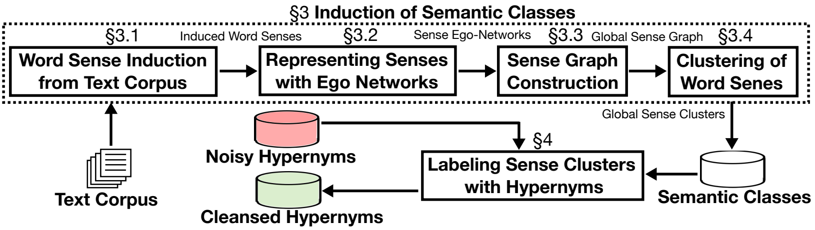
1. Sense-aware distributional semantic classes are **induced from a text corpus**;
2. Semantic classes are used to **filter a noisy hypernyms** database.

1. Sense-aware distributional semantic classes are **induced from a text corpus**;
2. Semantic classes are used to **filter a noisy hypernyms** database.

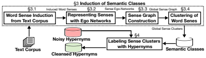
(e.g. extracted by an external method from a text corpus)

Outline of our approach

1. Sense-aware distributional semantic classes are **induced from a text corpus**;
2. Semantic classes are used to **filter a noisy hypernyms** database.



1. Sense-aware distributional semantic classes are **induced from a text corpus**;
2. Semantic classes are used to **filter a noisy hypernyms** database.



(e.g. extracted by an external method from a text corpus)

Chinese Whispers graph clustering

Used for word sense induction, used for global clustering ...

Sample of induced sense inventory

Word Sense	Local Sense Cluster: Related Senses	Hypernyms
mango#0	peach#1, grape#0, plum#0, apple#0, apricot#0, watermelon#1, banana#1, coconut#0, pear#0, fig#0, melon#0, **mangosteen#0**, ...	fruit#0, food#0, ...
apple#0	mango#0, pineapple#0, banana#1, melon#0, grape#0, peach#1, watermelon#1, apricot#0, cranberry#0, pumpkin#0, **mangosteen#0**, ...	fruit#0, crop#0, ...
Java#1	C#4, Python#3, Apache#3, Ruby#6, Flash#1, C++#0, SQL#0, ASP#2, Visual Basic#1, CSS#0, Delphi#2, MySQL#0, Excel#0, Pascal#0, ...	programming language#3, language#0, ...
Python#3	PHP#0, Pascal#0, Java#1, SQL#0, Visual Basic#1, C++#0, JavaScript#0, Apache#3, Haskell#5, .NET#1, C#4, SQL Server#0, ...	language#0, technology#0, ...

Word Sense	Local Sense Cluster: Related Senses	Hypernyms
mango#0	peach#1, grape#0, plum#0, apple#0, apricot#0, watermelon#1, banana#1, coconut#0, pear#0, fig#0, melon#0, **mangosteen#0**, ...	fruit#0, food#0, ...
apple#0	mango#0, pineapple#0, banana#1, melon#0, grape#0, peach#1, watermelon#1, apricot#0, cranberry#0, pumpkin#0, **mangosteen#0**, ...	fruit#0, crop#0, ...
Java#1	C#4, Python#3, Apache#3, Ruby#6, Flash#1, C++#0, SQL#0, ASP#2, Visual Basic#1, CSS#0, Delphi#2, MySQL#0, Excel#0, Pascal#0, ...	programming language#3, language#0, ...
Python#3	PHP#0, Pascal#0, Java#1, SQL#0, Visual Basic#1, C++#0, JavaScript#0, Apache#3, Haskell#5, .NET#1, C#4, SQL Server#0, ...	language#0, technology#0, ...

entries representing “fruits” and “programming language” senses. Each word sense s is represented with a list of related senses $\mathcal{N}(s)$ and the list of hypernyms $\mathcal{H}(s)$. The hypernyms can be used as human-interpretable sense labels of the sense clusters. One sense s, such as “apple#0”, can appear in multiple entries.

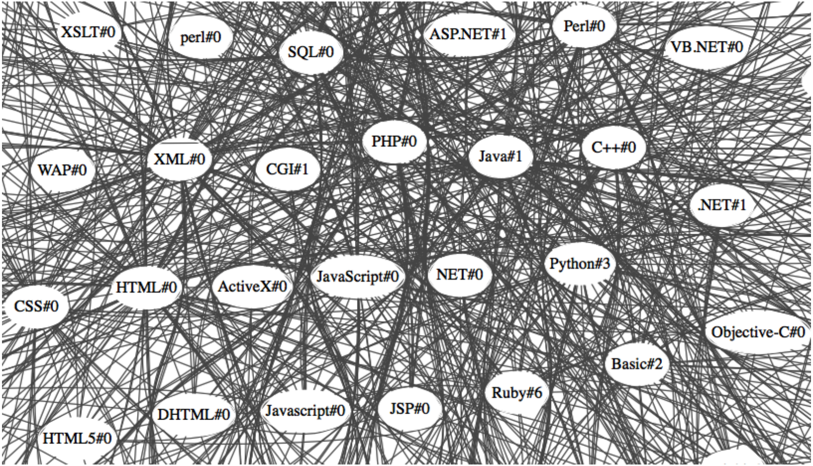
Sample of induced semantic classes

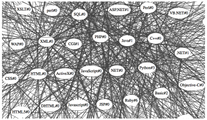
ID	Global Sense Cluster: Semantic Class	Hypernyms
1	peach#1, banana#1, pineapple#0, berry#0, blackberry#0, grapefruit#0, strawberry#0, blueberry#0, mango#0, grape#0, melon#0, orange#0, pear#0, plum#0, raspberry#0, watermelon#0, apple#0, apricot#0, watermelon#0, pumpkin#0, berry#0, **mangosteen#0**, ...	vegetable#0, fruit#0, crop#0, ingredient#0, food#0, ·
2	C#4, Basic#2, Haskell#5, Flash#1, Java#1, Pascal#0, Ruby#6, PHP#0, Ada#1, Oracle#3, Python#3, Apache#3, Visual Basic#1, ASP#2, Delphi#2, SQL Server#0, CSS#0, AJAX#0, JavaScript#0, SQL Server#0, Apache#3, Delphi#2, Haskell#5, .NET#1, CSS#0, ...	programming language#3, technology#0, language#0, format#2, app#0

ID	Global Sense Cluster: Semantic Class	Hypernyms
1	peach#1, banana#1, pineapple#0, berry#0, blackberry#0, grapefruit#0, strawberry#0, blueberry#0, mango#0, grape#0, melon#0, orange#0, pear#0, plum#0, raspberry#0, watermelon#0, apple#0, apricot#0, watermelon#0, pumpkin#0, berry#0, **mangosteen#0**, ...	vegetable#0, fruit#0, crop#0, ingredient#0, food#0, ·
2	C#4, Basic#2, Haskell#5, Flash#1, Java#1, Pascal#0, Ruby#6, PHP#0, Ada#1, Oracle#3, Python#3, Apache#3, Visual Basic#1, ASP#2, Delphi#2, SQL Server#0, CSS#0, AJAX#0, JavaScript#0, SQL Server#0, Apache#3, Delphi#2, Haskell#5, .NET#1, CSS#0, ...	programming language#3, technology#0, language#0, format#2, app#0

Sample of the induced sense clusters representing "fruits" and "programming language" semantic classes. Similarly to the induced word senses, the semantic classes are labeled with hypernyms. In contrast to the induced word senses, which represent a local clustering of word senses (related to a given word) semantic classes represent a global sense clustering of word senses. One sense *c*, such as "apple#0", can appear only in a single cluster.

Network of induced word senses





Senses referring to programming languages co-occur in global sense cluster entries, resulting in a densely connected set of co-hyponyms.

Optimization of meta-parameters

Comparison to WordNet and BabelNet

Meta-parameters

1. **Min. num. of sense co-occurrences in an ego-network**: $t > 0$
2. **Sense edge weight type**: count or log(count)
3. **Hypernym weight type**: tf-idf or tf

Meta-parameters

1. **Min. num. of sense co-occurrences in an ego-network**: $t > 0$
2. **Sense edge weight type**: count or log(count)
3. **Hypernym weight type**: tf-idf or tf

The method has really just a few parameters, but still we wanted to know their impact...

Since we are in an unsupervised setting...

Performance of different configurations of the hypernymy labeled global sense clusters in terms of their similarity to WordNet/BabelNet. The results are sorted by performance on BabelNet dataset, the best values in each section are boldfaced. The two underlined configurations are respectively the best *coarse-grained* and *fine-grained* grained semantic class models used in all experiments. The coarse grained model contains less semantic classes, but they tend to be more consistent than those of the fine-grained model, which contains more senses and classes.

Comparison to WordNet and BabelNet

Meta-parameters

1. **Min. num. of sense co-occurrences in an ego-network**: $t > 0$
2. **Sense edge weight type**: count or log(count)
3. **Hypernym weight type**: tf-idf or tf

$$hpc\text{-score}(c) = \frac{h\text{-score}(c) + 1}{p\text{-score}(c) + 1} \cdot \text{coverage}(c).$$

$$p\text{-score}(c) = \frac{1}{|c|} \sum_{i=1}^{|c|} \sum_{j=1}^{i} \text{dist}(w_i, w_j). \; h\text{-score}(c) = \frac{|\mathcal{H}(c) \cap \text{gold}(c)|}{|\mathcal{H}(c)|}.$$

Meta-parameters

1. **Min. num. of sense co-occurrences in an ego-network**: $t > 0$
2. **Sense edge weight type**: count or log(count)
3. **Hypernym weight type**: tf-idf or tf

$$hpc\text{-score}(c) = \frac{h\text{-score}(c) + 1}{p\text{-score}(c) + 1} \cdot \text{coverage}(c).$$

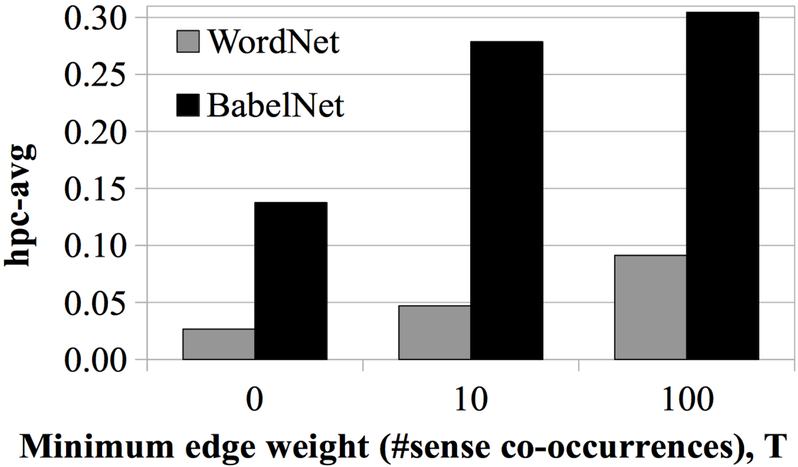
$$p\text{-score}(c) = \frac{1}{|c|} \sum_{i=1}^{|c|} \sum_{j=1}^{i} \text{dist}(w_i, w_j).\ h\text{-score}(c) = \frac{|\mathcal{H}(c) \cap \text{gold}(c)|}{|\mathcal{H}(c)|}.$$

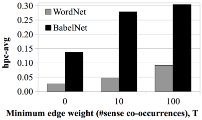
The method has really just a few parameters, but still we wanted to know their impact...

Since we are in an unsupervised setting...

Performance of different configurations of the hypernymy labeled global sense clusters in terms of their similarity to WordNet/BabelNet. The results are sorted by performance on BabelNet dataset, the best values in each section are boldfaced. The two underlined configurations are respectively the best *coarse-grained* and *fine-grained* grained semantic class models used in all experiments. The coarse grained model contains less semantic classes, but they tend to be more consistent than those of the fine-grained model, which contains more senses and classes.

Impact of the min. edge weight *t*





The higher the threshold the better quality relations we get, … but the smaller the network becomes

Best coarse- and fine-grained models

Min. num of sense co-occurr., *t*	**Edge weight**, *E*	**Hypernym weight**, *H*	**Number of clusters**	**Number of senses**	hpc-avg, **WordNet**	hpc-avg, **BabelNet**
0	count	tf-idf	**1 870**	**208 871**	0.041	0.279
100	log	tf-idf	**734**	**18 028**	0.092	0.304

Results

Plausibility of Semantic Classes

- Layout of the sense cluster evaluation crowdsourcing task;
- the entry "winchester" is the intruder.

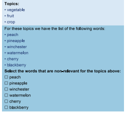
Topics:
- vegetable
- fruit
- crop

For these topics we have the list of the following words:
- peach
- pineapple
- winchester
- watermelon
- cherry
- blackberry

Select the words that are non-relevant for the topics above:

☐ peach
☐ pineapple
☐ winchester
☐ watermelon
☐ cherry
☐ blackberry

- Layout of the sense cluster evaluation crowdsourcing task;
- the entry “winchester” is the intruder.



Comparison to gold standard resources allows us to gauge the relative performances of various configurations of our method. To measure the absolute quality of the best configuration selected in the previous section, we rely on microtask-based crowdsourcing with CrowdFlower. We used two crowdsourcing tasks based on word intruder detection [Chang et al., 2009] to measure how humans perceive the extracted lexical-semantic structures. Namely, the tasks are designed to evaluate the quality of the extracted sense clusters and their labels. The input form presented to an annotator is illustrated in.
A crowdworker is asked to identify words that do not match the context represented by words from a sense cluster or its label. To generate an intruder, following the original design of , we select a random word from a cluster and replace it with a word of similar frequency that does not belong to any cluster (bias here is low as the evaluated model contains 27,149 out of 313,841 induced word senses). In both tasks, the workers have been provided with concise instructions and test questions.

Plausibility of Semantic Classes

1. **Accuracy** is the fraction of tasks where annotators correctly identified the intruder;
2. **Badness**: is the fraction of tasks for which non-intruder words were selected.

1. **Accuracy** is the fraction of tasks where annotators correctly identified the intruder;
2. **Badness**: is the fraction of tasks for which non-intruder words were selected.

Plausibility of the sense clusters according to human judgments via an intruder detection experiment for the coarse-grained semantic class model.

Plausibility of Semantic Classes

1. **Accuracy** is the fraction of tasks where annotators correctly identified the intruder;
2. **Badness**: is the fraction of tasks for which non-intruder words were selected.

	Accuracy	**Badness**	**Randolph κ**
Sense clusters, c	0.859	0.248	0.739
Hyper. labels, $\mathcal{H}(c)$	0.919	0.208	0.705

1. **Accuracy** is the fraction of tasks where annotators correctly identified the intruder;
2. **Badness**: is the fraction of tasks for which non-intruder words were selected.

	Accuracy	Badness	Randolph κ
Sense clusters, c	0.859	0.248	0.739
Hyper. labels, $\mathcal{H}(c)$	0.919	0.208	0.705

Plausibility of the sense clusters according to human judgments via an intruder detection experiment for the coarse-grained semantic class model.

Plausibility of Semantic Classes

1. **Accuracy** is the fraction of tasks where annotators correctly identified the intruder;
2. **Badness**: is the fraction of tasks for which non-intruder words were selected.

	Accuracy	**Badness**	**Randolph κ**
Sense clusters, c	0.859	0.248	0.739
Hyper. labels, $\mathcal{H}(c)$	0.919	0.208	0.705

- **Clusters**: 68 annotators, 2,035 judgments;
- **Hypernyms**: 98 annotators, 2,245 judgments.

1. **Accuracy** is the fraction of tasks where annotators correctly identified the intruder;
2. **Badness**: is the fraction of tasks for which non-intruder words were selected.

	Accuracy	Badness	Randolph κ
Sense clusters, c	0.859	0.248	0.739
Hyper. labels, $\mathcal{H}(c)$	0.919	0.208	0.705

- **Clusters**: 68 annotators, 2,035 judgments;
- **Hypernyms**: 98 annotators, 2,245 judgments.

Plausibility of the sense clusters according to human judgments via an intruder detection experiment for the coarse-grained semantic class model.

Improving Binary Hypernymy Relations

Layout of the hypernymy annotation task:

Is it correct that **peach** is a kind of **fruit**?

Your opinion:

- ○ Yes
- ○ No

Improving Binary Hypernymy Relations

Evaluating results of post-processing of a noisy hypernymy database using human judgements:

- A random sample of 4,870 relations using lexical split;
- each labeled 6.9 times on average;
- a total of 33,719 judgments.

Improving Binary Hypernymy Relations

Evaluating results of post-processing of a noisy hypernymy database using human judgements:

- A random sample of 4,870 relations using lexical split;
- each labeled 6.9 times on average;
- a total of 33,719 judgments.

	Precision	**Recall**	**F-score**
Original hypernymy relations extracted from Common Crawl corpus [Seitner et al., 2016]	0.475	0.546	0.508
Enhanced hypernyms with the *coarse-grained* semantic classes	**0.541**	**0.679**	**0.602**

Improving Taxonomy Induction

- **SemEval 2016 Task 13** ”Taxonomy Extraction from Text”;
- **Fowlkes&Mallows Measure (F&M)** – a cumulative measure of the similarity of taxonomies;
- **English** part of the dataset.

- **SemEval 2016 Task 13** ”Taxonomy Extraction from Text”;
- **Fowlkes&Mallows Measure (F&M)** – a cumulative measure of the similarity of taxonomies;
- **English** part of the dataset.

Summary of the domain-specific sense clusters.

Improving Taxonomy Induction

- **SemEval 2016 Task 13** ”Taxonomy Extraction from Text”;
- **Fowlkes&Mallows Measure (F&M)** – a cumulative measure of the similarity of taxonomies;
- **English** part of the dataset.

Domain	**#Seeds words**	**#Expanded words**	**#Clusters,** fine-gr.	**#Clusters,** coarse-gr.
Food	2 834	3 047	29	21
Science	806	1 137	73	35
Environ.	261	909	111	39

- **SemEval 2016 Task 13** "Taxonomy Extraction from Text";
- **Fowlkes&Mallows Measure (F&M)** – a cumulative measure of the similarity of taxonomies;
- **English** part of the dataset.

Domain	**#Seeds words**	**#Expanded words**	**#Clusters,** fine-gr.	**#Clusters,** coarse-gr.
Food	2 834	3 047	29	21
Science	806	1 137	73	35
Environ.	261	909	111	39

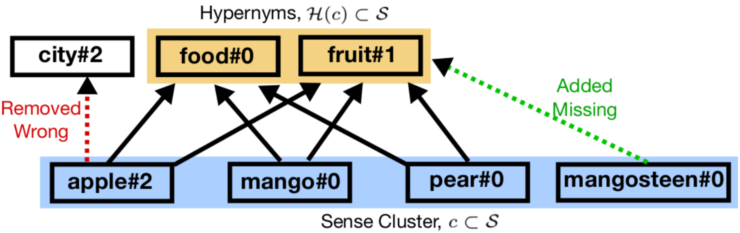
Summary of the domain-specific sense clusters.

Improving Taxonomy Induction

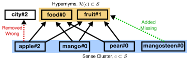
System / Dataset	Food, WordNet	Science, WordNet	Food, Combined	Science, Combined	Science, Eurovoc	Environ., Eurovoc
WordNet	1.0000	1.0000	0.5870	0.5760	0.6243	n.a.
Baseline	0.0022	0.0016	0.0019	0.0163	0.0056	0.0000
JUNLP	0.1925	0.0494	0.2608	0.1774	0.1373	0.0814
NUIG-UNLP	n.a.	0.0027	n.a.	0.0090	0.1517	0.0007
QASSIT	n.a.	0.2255	n.a.	0.5757	0.3893	0.4349
TAXI	0.3260	0.2255	0.2021	0.3634	0.3893	0.2384
USAAR	0.0021	0.0008	0.0000	0.0020	0.0023	0.0007
Sem. Class, fine-gr.	0.4540	0.4181	0.5147	0.6359	**0.5831**	0.5600
Sem. Class, coarse-gr.	**0.4774**	**0.5927**	**0.5799**	**0.6539**	0.5515	**0.6326**

Summary

1. An unsupervised method for the induction of **sense-aware distributional semantic classes**;
2. Showed how these can be used for **post-processing of noisy hypernymy databases** extracted from text.



1. An unsupervised method for the induction of **sense-aware distributional semantic classes**;
2. Showed how these can be used for **post-processing of noisy hypernymy databases** extracted from text.



By using global as opposed to local information ...

Thank you! Questions?

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009).
Reading Tea Leaves: How Humans Interpret Topic Models.
In *Advances in Neural Information Processing Systems 22* (pp. 288–296). Curran Associates, Inc.

Gong, Z., Cheang, C. W., & Leong Hou, U. (2005).
Web Query Expansion by WordNet.
In *Proceedings of the 16th International Conference on Database and Expert Systems Applications - DEXA '05* (pp. 166–175). Copenhagen, Denmark: Springer Berlin Heidelberg.

Seitner, J., Bizer, C., Eckert, K., Faralli, S., Meusel, R., Paulheim, H., & Ponzetto, S. P. (2016).
A Large DataBase of Hypernymy Relations Extracted from the Web.
In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016 (pp. 360–367). Portorož, Slovenia: European Language Resources Association (ELRA).

Shi, L. & Mihalcea, R. (2005).
Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing.
In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing 2005 (pp. 100–111). Mexico City, Mexico: Springer Berlin Heidelberg.

Zhou, G., Liu, Y., Liu, F., Zeng, D., & Zhao, J. (2013).
Improving question retrieval in community question answering using world knowledge.
In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13 (pp. 2239–2245). Beijing, China: AAAI Press.