Unsupervised discovery of negative categories in lexicon bootstrapping

Tara McIntosh
NICTA Victoria Research Lab
University of Melbourne
nlp@taramcintosh.org

Abstract

Multi-category bootstrapping algorithms were developed to reduce semantic drift. By extracting multiple semantic lexicons simultaneously, a category's search space may be restricted. The best results have been achieved through reliance on manually crafted negative categories. Unfortunately, identifying these categories is non-trivial, and their use shifts the unsupervised bootstrapping paradigm towards a supervised framework.

We present NEG-FINDER, the first approach for discovering negative categories automatically. NEG-FINDER exploits unsupervised term clustering to generate multiple negative categories during bootstrapping. Our algorithm effectively removes the necessity of manual intervention and formulation of negative categories, with performance closely approaching that obtained using negative categories defined by a domain expert.

1 Introduction

Automatically acquiring semantic lexicons from text is essential for overcoming the knowledge bottleneck in many NLP tasks, e.g. question answering (Ravichandran and Hovy, 2002). Many of the successful methods follow the unsupervised iterative bootstrapping framework (Riloff and Shepherd, 1997). Bootstrapping has since been effectively applied to extracting general semantic lexicons (Riloff and Jones, 1999), biomedical entities (Yu and Agichtein, 2003) and facts (Carlson et al., 2010).

Bootstrapping is often considered to be minimally supervised, as it is initialised with a small set of seed terms of the target category to extract. These seeds are used to identify patterns that can match the target category, which in turn can extract new lexicon terms (Riloff and Jones, 1999). Unfortunately, *semantic drift* often occurs when ambiguous or erroneous terms and/or patterns are introduced into the iterative process (Curran et al., 2007).

In multi-category bootstrapping, semantic drift is often reduced when the target categories compete with each other for terms and/or patterns (Yangarber et al., 2002). This process is most effective when the categories bound each other's search space. To ensure this, manually crafted negative categories are introduced (Lin et al., 2003; Curran et al., 2007). Unfortunately, this makes these algorithms substantially more supervised.

The design of negative categories is a very time consuming task. It typically requires a domain expert to identify the semantic drift and its cause, followed by a significant amount of trial and error in order to select the most suitable combination of negative categories. This introduces a substantial amount of supervised information into what was an unsupervised framework, and in turn negates one of the main advantages of bootstrapping — the quick construction of accurate semantic lexicons.

We show that although excellent performance is achieved using negative categories, it varies greatly depending on the negative categories selected. This highlights the difficulty of crafting negative categories and thus the necessity for tools that can automatically identify them.

Our second contribution is the first fully unsupervised approach, NEG-FINDER, for discovering

negative categories automatically. During bootstrapping, efficient clustering techniques are applied to sets of drifted candidate terms to generate new negative categories. Once a negative category is identified it is incorporated into the subsequent iterations whereby it provides the necessary semantic boundaries for the target categories.

We demonstrate the effectiveness of our approach for extracting biomedical semantic lexicons by incorporating NEG-FINDER within the WMEB-DRIFT bootstrapping algorithm (McIntosh and Curran, 2009). NEG-FINDER significantly outperforms bootstrapping prior to the domain expert's negative categories. We show that by using our discovered categories we can reach near expert-guided performance. Our methods effectively remove the necessity of manual intervention and formulation of negative categories in semantic lexicon bootstrapping.

2 Background

Various automated pattern-based bootstrapping algorithms have been proposed to iteratively build semantic lexicons. In multi-level bootstrapping, a lexicon is iteratively expanded from a small sample of seed terms (Riloff and Jones, 1999). The seed terms are used to identify contextual patterns they appear in, which in turn may be used to extract new lexicon entries. This process is repeated with the new expanded lexicon identifying new patterns.

When bootstrapping semantic lexicons, polysemous or erroneous terms and/or patterns that weakly constrain the semantic class are eventually extracted. This often causes *semantic drift* — when a lexicon's intended meaning shifts into another category during bootstrapping (Curran et al., 2007). For example, *female names* may drift into *gemstones* when the terms *Ruby* and *Pearl* are extracted.

Multi-category bootstrapping algorithms, such as BASILISK (Thelen and Riloff, 2002), NOMEN (Yangarber et al., 2002), and WMEB (McIntosh and Curran, 2008), aim to reduce semantic drift by extracting multiple semantic categories simultaneously. These algorithms utilise information about other semantic categories in order to reduce the categories from drifting towards each other. This framework has recently been extended to extract different relations from text (Carlson et al., 2010).

2.1 Weighted MEB

In Weighted Mutual Exclusion Bootstrapping (WMEB, McIntosh and Curran, 2008), multiple semantic categories iterate simultaneously between the term and pattern extraction phases, competing with each other for terms and patterns. Semantic drift is reduced by forcing the categories to be mutually exclusive. That is, candidate terms can only be extracted by a single category and patterns can only extract terms for a single category.

In WMEB, multiple bootstrapping instances are initiated for each competing target category. Each category's seed set forms its initial lexicon. For each term in the category lexicon, WMEB identifies all candidate contextual patterns that can match the term in the text. To ensure mutual exclusion between the categories, candidate patterns that are identified by multiple categories in an iteration are excluded. The remaining patterns are then ranked according to the *reliability measure* and *relevance weight*.

The reliability of a pattern for a given category is the number of extracted terms in the category's lexicon that match the pattern. A pattern's relevance weight is defined as the sum of the χ-squared values between the pattern (p) and each of the lexicon terms (t): $\text{weight}(p) = \sum_{t \in T} \chi^2(p, t)$. These metrics are symmetrical for both candidate terms and patterns.

The top-m patterns are then added to the pool of extracting patterns. If each of the top-m patterns already exists in the pool, the next unseen pattern is added to the pool. This ensures at least one new pattern is added to the pool in each iteration.

In the term selection phase, all patterns within the pattern pool are used to identify candidate terms. Like the candidate patterns, terms that are extracted by multiple categories in the same iteration are also excluded. The remaining candidate terms are ranked with respect to their reliability and relevance weight, and the top-n terms are added to the lexicon.

2.2 Detecting semantic drift in WMEB

In McIntosh and Curran (2009), we showed that multi-category bootstrappers are still prone to semantic drift in the later iterations. We proposed a drift detection metric based on our hypothesis that semantic drift occurs when a candidate term is more similar to the recently added terms than to the seed

and high precision terms extracted in the earlier iterations. Our metric is based on distributional similarity measurements and can be directly incorporated into WMEB's term selection phase to prevent drifting terms from being extracted (WMEB-DRIFT).

The drift metric is defined as the ratio of the average distributional similarity of the candidate term to the first n terms extracted into the lexicon L, and to the last m terms extracted in the previous iterations:

$$\text{drift}(term, n, m) = \frac{\text{avgsim}(L_{1...n}, term)}{\text{avgsim}(L_{(N-m+1)...N}, term)} \quad (1)$$

2.3 Negative categories

In multi-category bootstrapping, improvements in precision arise when semantic boundaries between multiple target categories are established. Thus, it is beneficial to bootstrap categories that share similar semantic spaces, such as *female names* and *flowers*.

Unfortunately, it is difficult to predict if a target category will suffer from semantic drift and/or whether it will naturally compete with the other target categories. Once a domain expert establishes semantic drift and its possible cause, a set of *negative/stop* categories that may be of no direct interest are manually crafted to prevent semantic drift. These additional categories are then exploited during another round of bootstrapping to provide further competition for the target categories (Lin et al., 2003; Curran et al., 2007).

Lin et al. (2003) improved NOMEN's performance for extracting *diseases* and *locations* from the ProMED corpus by incorporating negative categories into the bootstrapping process. They first used one general negative category, seeded with the 10 most frequent nouns in the corpus that were unrelated to the target categories. This single negative category resulted in substantial improvements in precision. In their final experiment, six negative categories that were notable sources of semantic drift were identified, and the inclusion of these lead to further performance improvements ($\sim$20%).

In similar experiments, both Curran et al. (2007) and McIntosh (2010) manually crafted negative categories that were necessary to prevent semantic drift. In particular, in McIntosh (2010), a biomedical expert spent considerable time ($\sim$15 days) and effort identifying potential negative categories and subsequently optimising their associated seeds in trial and error bootstrapping runs.

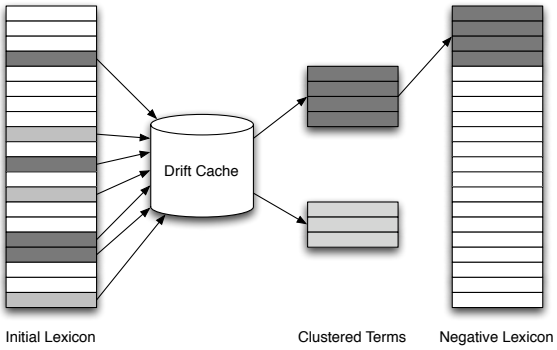

Figure 1: NEG-FINDER: Local negative discovery

By introducing manually crafted negative categories, a significant amount of expert domain knowledge is introduced. The use of this expert knowledge undermines the principle advantages of unsupervised bootstrapping, by making it difficult to bootstrap lexicons for a large number of categories across diverse domains or languages. In this paper, we aim to push multi-category bootstrapping back into its original minimally-supervised framework, with as little performance loss as possible.

3 NEG-FINDER

Our approach, *Negative Category Finder for Bootstrapping* (NEG-FINDER), can be easily incorporated into bootstrapping algorithms that exclude candidate terms or facts based on a selection criteria, including WMEB-DRIFT and Paşca et al.'s (2006) large-scale fact extraction system. For simplicity, we describe our approach within the WMEB-DRIFT bootstrapping algorithm. Figure 1 shows the framework of our approach.

To discover negative categories during bootstrapping, NEG-FINDER must identify a representative cluster of the drifted terms. In this section, we present the two types of clustering used (*maximum* and *outlier*), and our three different levels of negative discovery (*local*, *global* and *mixture*).

3.1 Discovering negative categories

We have observed that semantic drift begins to dominate when clusters of incorrect terms with similar

meanings are extracted. In the term selection phase of WMEB-DRIFT, the top-n candidate terms that satisfy the drift detection threshold are added to the expanding lexicon. Those terms which are considered but do not meet the threshold are excluded.

In NEG-FINDER, these drifted terms are cached as they may provide adequate seed terms for new negative categories. However, the drifted terms can also include scattered polysemous or correct terms that share little similarity with the other drifted terms. Therefore, simply using the first set of drifted terms to establish a negative category is likely to introduce noise rather than a cohesive competing category.

To discover negative categories, we exploit hierarchical clustering to group similar terms within the cache of drifted terms. In agglomerative hierarchical clustering, a single term is assigned to an individual cluster, and these clusters are iteratively merged until a final cluster is formed containing all terms (Kaufmann and Rousseeuw, 1990). In our approach, the similarity between two clusters is computed as the average distributional similarity between all pairs of terms across the clusters (average-link clustering).

For calculating the similarity between two terms we use the distributional similarity approach described in Curran (2004). We extracted window-based features from the set of candidate patterns to form context vectors for each term. We use the standard t-test weight and weighted Jaccard measure functions (Curran, 2004).

To ensure adequate coverage of the possible drifting topics, negative discovery and hence clustering is only performed when the drift cache consists of at least 20 terms.

3.2 Maximum and outlier clustering

Although hierarchical clustering is quadratic, we can efficiently exploit the agglomerative process as the most similar terms will merge into clusters first. Therefore, to identify the k most similar terms, we can exit the clustering process as soon as a cluster of size k is established. We refer to this approach as *maximum clustering*.

In our next clustering method, we aim to form a negative category with as little similarity to the target seeds. We use an *outlier clustering* strategy, in which the drifted term t with the least average distributional similarity to the first n terms in the lexicon must be contained in the cluster of seeds. We use average similarity to the first n terms, as it is already pre-computed for the drift detection metric. As with *maximum clustering*, once a cluster of size k containing the term t is formed, the clustering process can be terminated.

3.3 Incorporating the negative category

After a cluster of negative seed terms is established, the drift cache is cleared, and a new negative category is created and introduced into the iterative bootstrapping process in the next iteration. This means that the negative category can only influence the subsequent iterations of bootstrapping. The negative categories can compete with all other categories, including any previously introduced negative categories, however the negative categories do not contribute to the drift caches.

Before the new category is introduced, its first set of extracting patterns must be identified. For this, the complete set of extracting patterns matching any of the negative seeds are considered and ranked with respect to the seeds. The top scoring patterns are considered sequentially until m patterns are assigned to the new negative category. To ensure mutual exclusion between the new category and the target categories, a candidate pattern that has previously been selected by a target category cannot be used to extract terms for either category in the subsequent iterations.

3.4 Levels of negative discovery

Negative category discovery can be performed at a *local* or *global* level, or as a *mixture* of both. In *local discovery*, each target category has its own drifted term cache and can generate negative categories irrespective of the other target categories. This is shown in Figure 1. The drifted terms (shaded) are extracted away from the lexicon into the local drift cache, which is then clustered. A cluster is then used to initiate a negative category's lexicon. Target categories can also generate multiple negative categories across different iterations.

In *global discovery*, all drifted terms are pooled into a global cache, from which a single negative category can be identified in an iteration. This is based on our intuition that multiple target categories

TYPE	MEDLINE
No. Terms	1 347 002
No. Patterns	4 090 412
No. 5-grams	72 796 760
No. Unfiltered tokens	6 642 802 776

Table 1: Filtered 5-gram dataset statistics.

may be drifting into similar semantic categories, and enables these otherwise missed negative categories to be established.

In the *mixture discovery* method, both global and local negative categories can be formed. A category's drifted terms are collected into its local cache as well as the global cache. Negative discovery is then performed on each cache when they contain at least 20 terms. Once a local negative category is formed, the terms within the local cache are cleared and also removed from the global cache. This prevents multiple negative categories being instantiated with overlapping seed terms.

4 Experimental setup

To compare the effectiveness of our negative discovery approaches we consider the task of extracting biomedical semantic lexicons from raw text.

4.1 Data

The algorithms take as input a set of candidate terms to be extracted into semantic lexicons. The source text collection consists of 5-grams (t_1, t_2, t_3, t_4, t_5) from approximately 16 million MEDLINE abstracts.[1] The set of possible candidate terms correspond to the middle tokens (t_3), and the possible patterns are formed from the surrounding tokens (t_1, t_2, t_4, t_5). We do not use syntactic knowledge, as we did not wish to rely on any tools that require supervised training, to ensure our technique is as domain and language independent as possible.

Limited preprocessing was required to extract the 5-grams from MEDLINE. The XML markup was removed, and the collection was tokenised and split into sentences using bio-specific NLP tools (Grover et al., 2006). Filtering was applied to remove infrequent patterns and terms – patterns appearing with less than 7 different terms, and terms only appearing with those patterns were removed. The statistics of the resulting dataset are shown in Table 1.

[1]The set contains all MEDLINE titles and abstracts available up to Oct 2007.

CAT	DESCRIPTION
ANTI	Antibodies: *MAb IgG IgM rituximab infliximab* (κ_1:0.89, κ_2:1.0)
CELL	Cells: *RBC HUVEC BAEC VSMC SMC* (κ_1:0.91, κ_2:1.0)
CLNE	Cell lines: *PC12 CHO HeLa Jurkat COS* (κ_1:0.93, κ_2: 1.0)
DISE	Diseases: *asthma hepatitis tuberculosis HIV malaria* (κ_1:0.98, κ_2:1.0)
DRUG	Drugs: *acetylcholine carbachol heparin penicillin tetracyclin* (κ_1:0.86, κ_2:0.99)
FUNC	Molecular functions and processes: *kinase ligase acetyltransferase helicase binding* (κ_1:0.87, κ_2:0.99)
MUTN	Protein and gene mutations: *Leiden C677T C282Y 35delG null* (κ_1:0.89, κ_2:1.0)
PROT	Proteins and genes: *p53 actin collagen albumin IL-6* (κ_1:0.99, κ_2:1.0)
SIGN	Signs and symptoms: *anemia fever hypertension hyperglycemia cough* (κ_1:0.96, κ_2:0.99)
TUMR	Tumors: *lymphoma sarcoma melanoma osteosarcoma neuroblastoma* (κ_1:0.89, κ_2:0.95)

Table 2: The MEDLINE semantic categories

4.2 Semantic categories

The semantic categories we extract from MEDLINE were inspired by the TREC Genomics entities (Hersh et al., 2007) and are described in detail in McIntosh (2010). The hand-picked seeds selected by a domain expert for each category are shown in italics in Table 2. These were carefully chosen to be as unambiguous as possible with respect to the other categories.

4.3 Negative categories

In our experiments, we use two different sets of negative categories. These are shown in Table 3. The first set corresponds to those used in McIntosh and Curran (2008), and were identified by a domain expert as common sources of semantic drift in preliminary experiments with MEB and WMEB. The AMINO ACID category was created in order to filter common MUTN errors. The ANIMAL and BODY PART categories were formed with the intention of preventing drift in the CELL, DISE and SIGN categories. The ORGANISM category was then created to reduce the new drift forming in the DISE category after the first set of negative categories were introduced.

The second set of negative categories was identified by an independent domain expert with limited

	CATEGORY	SEED TERMS
1	AMINO ACID	arginine cysteine glycine glutamate histamine
	ANIMAL	insect mammal mice mouse rats
	BODY PART	breast eye liver muscle spleen
	ORGANISM	Bartonella Borrelia Cryptosporidium Salmonella toxoplasma
2	AMINO ACID	Asn Gly His Leu Valine
	ANIMAL	animals dogs larvae rabbits rodents
	ORGANISM	Canidia Shigella Scedosporium Salmonella Yersinia
	GENERIC	decrease effects events increase response
	MODIFIERS	acute deep intrauterine postoperative secondary
	PEOPLE	children females men subjects women
	SAMPLE	biopsies CFU sample specimens tissues

Table 3: Manually crafted negative categories

knowledge of NLP and bootstrapping. This expert identified three similar categories to the first expert, however their seeds are very different. They also identified three more categories than the first.

4.4 Lexicon evaluation

Our evaluation process follows that of McIntosh and Curran (2009) and involved manually inspecting each extracted term and judging whether it was a member of the semantic class. This manual evaluation was performed by two domain experts and is necessary due to the limited coverage of biomedical resources. Inter-annotator agreement scores are provided in Table 2.[2] To make later evaluations more efficient, all evaluators' decisions for each category are cached.

Unfamiliar terms were checked using online resources including MEDLINE, MeSH, and Wikipedia. Each ambiguous term was counted as correct if it was classified into one of its correct categories, such as *lymphoma*, which is a TUMR and DISE. If a term was unambiguously part of a multi-word term we considered it correct. Abbreviations, acronyms, and obvious misspelled words were included.

For comparing the performance of the algorithms, the average precision for the top-1000 terms over the 10 target categories is measured. To identify when semantic drift has a significant impact, we report the precision of specific sections of the lexicon, e.g. the 801-1000 sample corresponds to the last 200 terms.

[2]All disagreements were discussed, and the kappa scores κ_1 and κ_2 are those before and after the discussions, respectively.

	1-500	1-1000
WMEB-DRIFT	74.3	68.6
+negative 1	87.7	82.8
+negative 2	83.8	77.8

Table 4: Influence of negative categories

4.5 System settings

All experiments were performed using the 10 target categories as input. Unless otherwise stated, no hand-picked negative categories are used.

Each target category is initialised with the 5 hand-picked seed terms (Table 2). In each iteration a maximum of 5 lexicon terms and 5 new patterns can be extracted by a category. The bootstrapping algorithms are run for 200 iterations.

The drift detection metric is calculated over the first 100 terms and previous 5 terms extracted into the lexicon, and the filter threshold is set to 0.2, as in McIntosh and Curran (2009). To ensure infrequent terms are not used to seed negative categories, drifted terms must occur at least 50 times to be retained in the drift cache. Negative category discovery is only initiated when the drifted cache contains at least 20 terms, and a minimum of 5 terms are used to seed a negative category.

4.6 Random seed experiments

Both McIntosh and Curran (2009) and Pantel et al. (2009) have shown that a bootstrapper's performance can vary greatly depending on the input seeds. To ensure our methods are compared reliably, we also report the average precision of randomised seed experiments. Each algorithm is instantiated 10 times with different random gold seeds for each target category. These gold seeds are randomly sampled from the evaluation cache formed in McIntosh and Curran (2009).

5 Results

5.1 Influence of negative categories

In our first experiments, we investigate the performance variations and improvements gained using negative categories selected by two independent domain experts. Table 4 shows WMEB-DRIFT's average precision over the 10 target categories with and without the two negative category sets. Both

	1-200	201-400	401-600	601-800	801-1000	1-1000
WMEB-DRIFT	79.5	74.8	64.7	61.9	62.1	68.6
NEG-FINDER						
First discovered	79.5	74.3	64.8	67.8	66.6	70.7
Local discovery						
+maximum	79.5	74.8	67.3	69.3	70.5	72.2
+outlier	79.5	73.9	64.8	67.8	71.0	71.5
Global discovery						
+maximum	79.5	73.9	65.7	73.2	72.7	73.4
+outlier	79.5	74.7	65.6	71.4	68.2	72.1
Mixture discovery						
+maximum	79.5	74.7	69.3	73.3	72.8	74.0
+outlier	79.5	75.2	69.7	72.0	69.4	73.2

Table 5: Performance comparison of WMEB-DRIFT and NEG-FINDER

sets significantly improve WMEB-DRIFT, however there is a significant performance difference between them. This demonstrates the difficulty of selecting appropriate negative categories and seeds for the task, and in turn the necessity for tools to discover them automatically.

5.2 Negative category discovery

Table 5 compares the performance of NEG-FINDER incorporated with WMEB-DRIFT. Each method has equal average precision over the first 200 terms, as semantic drift does not typically occur in the early iterations. Each discovery method significantly outperforms WMEB-DRIFT in the later stages, and over the top 1000 terms.[3]

The *first discovery* approach corresponds to the naïve NEG-FINDER system that generates *local* negative categories from the first five drifted terms. Although it outperforms WMEB-DRIFT, its advantage is smaller than the clustering methods.

The *outlier clustering* approach, which we predicted to be the most effective, was surprisingly less accurate than the *maximum* approach for selecting negative seeds. This is because the seed cluster formed around the outlier term is not guaranteed to have high pair-wise similarity and thus it may represent multiple semantic categories.

Local discovery was the least effective discovery approach. Compared to local discovery, *global discovery* is capable of detecting new negative categories earlier, and the categories it detects are more likely to compete with multiple target categories.

[3]Statistical significance was tested using computationally-intensive randomisation tests (Cohen, 1995).

CATEGORY	NEGATIVE SEEDS
CELL-NEG	animals *After* Lambs Pigs Rabbits
TUMR-NEG	inoperable multinodular nonresectable operated unruptured
GLOBAL	days Hz mM post Torr
GLOBAL	aortas eyes legs mucosa retinas
GLOBAL	men offspring parents persons relatives
GLOBAL	Australian Belgian Dutch European Italian
GLOBAL	Amblyospora Branhamella Phormodium Pseudanabaena Rhodotorula

Table 6: Negative categories from mixture discovery

The NEG-FINDER *mixture* approach, which benefits from both *local* and *global discovery*, identifies the most useful negative categories. Table 6 shows the seven discovered categories — two local negative categories from CELL and TUMOUR, and five global categories were formed. Many of these categories are similar to those identified by the domain experts. For example, clear categories for ANIMAL, BODY PART, PEOPLE and ORGANISM are created. By identifying and then including these negative categories, NEG-FINDER significantly outperforms WMEB-DRIFT by 5.4% over the top-1000 terms and by 10.7% over the last 200 terms, where semantic drift is prominent. These results demonstrate that suitable negative categories can be identified and exploited during bootstrapping.

5.3 Boosting hand-picked negative categories

In our next set of experiments, we investigate whether NEG-FINDER can improve state-of-the-art performance by identifying new negative categories in addition to the manually selected negative

	1-200	201-400	401-600	601-800	801-1000	1-1000
WMEB-DRIFT						
+negative 1	90.5	87.3	82.0	74.6	79.8	82.8
+negative 2	87.8	82.2	78.7	76.1	63.3	77.8
WMEB-DRIFT						
+restart +local	85.5	82.6	76.5	75.7	68.5	78.4
+restart +global	84.0	83.8	79.1	74.8	69.5	79.7
+restart +mixture	85.2	85.0	82.3	72.5	72.7	81.4

Table 7: Performance of WMEB-DRIFT using negative categories discovered by NEG-FINDER

	601-800	801-1000	1-1000
WMEB-DRIFT			
+negative 1	74.6	79.8	82.8
NEG-FINDER			
+negative 1 +local	76.4	80.1	83.2
+negative 1 +global	77.5	76.0	82.7
+negative 1 +mixture	76.7	79.9	83.2

Table 8: Performance of NEG-FINDER with manually crafted negative categories

categories. Both NEG-FINDER and WMEB-DRIFT are initialised with the 10 target categories and the first set of negative categories.

Table 8 compares our best performing systems (NEG-FINDER *maximum clustering*) with standard WMEB-DRIFT, over the last 400 terms where semantic drift dominates. NEG-FINDER effectively discovers additional categories and significantly outperforms WMEB-DRIFT. This further demonstrates the utility of our approach.

5.4 Restarting with new negative categories

The performance improvements so far using NEG-FINDER have been limited by the time at which new negative categories are discovered and incorporated into the bootstrapping process. That is, system improvements can only be gained from the negative categories after they are generated. For example, in *Local* NEG-FINDER, five negative categories are discovered in iterations 83, 85, 126, 130 and 150. On the other hand, in the WMEB-DRIFT +negative experiments (Table 8 row 2), the hand-picked negative categories can start competing with the target categories in the very first iteration of bootstrapping.

To test the full utility of NEG-FINDER, we use the set of discovered categories as competing input for WMEB-DRIFT. Table 7 shows the average precision of WMEB-DRIFT over the 10 target categories when it is restarted with the new negative categories discovered from our three approaches (using *maximum clustering*). Over the first 200 terms, significant improvements are gained using the new negative categories (+6%). However, the manually selected categories are far superior in preventing drift (+11%). This may be attributed by the target categories not strongly drifting into the new negative categories until the later stages, whereas the hand-picked categories were selected on the basis of observed drift in the early stages (over the first 500 terms).

Each NEG-FINDER approach significantly outperforms WMEB-DRIFT with no negative categories. For example, using the NEG-FINDER *mixture* categories increases precision by 12.8%. These approaches also outperform their corresponding inline discovery methods (e.g. +7.4% with *mixture discovery* – Table 5).

Table 7 shows that each of the discovered negative sets can significantly outperform the negative categories selected by a domain expert (negative set 2) (+0.6 – 3.9%). Our best system's performance (*mixture*: 81.4%) closely approaches that of the superior negative set, trailing by only 1.4%.

5.5 Individual categories

In this section, we analyse the effect of NEG-FINDER on the individual target categories. Table 9 shows the average precision of the lexicons for some target categories. All categories, except TUMOUR, improve significantly with the inclusion of the discovered negative categories. In particular, the CELL and SIGN categories, which are affected severely by semantic drift, increase by up to 33.3% and 45.2%, respectively. The discovered negative categories are more effective than the manually crafted sets in reducing semantic drift in the ANTIBODY, CELL and DISEASE lexicons.

	ANTI	CELL	DISE	SIGN	TUMR
WMEB-DRIFT	92.9	47.8	49.3	27.9	39.5
+negative 1	91.6	73.1	87.8	76.5	48.7
+negative 2	85.8	68.0	84.2	71.3	16.3
NEG-FINDER					
+mixture	94.9	73.9	56.0	41.0	42.2
+mixture +negative 1	90.8	77.2	87.8	78.2	48.2
WMEB-DRIFT					
+restart +local	89.9	78.8	71.6	73.1	32.2
+restart +global	94.6	79.0	81.9	62.6	35.2
+restart +mixture	92.6	81.1	91.1	63.6	47.5

Table 9: Individual category results (1-1000 terms)

5.6 Random seed experiments

In Table 10, we report the results of our randomised experiments. Over the last 200 terms, WMEB-DRIFT with the first set of negative categories (row 2) is outperformed by NEG-FINDER (row 4). NEG-FINDER also significantly boosts the performance of the original negative categories by identifying additional negative categories (row 5). Our final experiment, where WMEB-DRIFT is re-initialised with the negative categories discovered by NEG-FINDER, further demonstrates the utility of our method. On average, the discovered negative categories significantly outperform the manually crafted negative categories.

6 Conclusion

In this paper, we have proposed the first completely unsupervised approach to identifying the negative categories that are necessary for bootstrapping large yet precise semantic lexicons. Prior to this work, negative categories were manually crafted by a domain expert, undermining the advantages of an unsupervised bootstrapping paradigm.

There are numerous avenues for further examination. We intend to use sophisticated clustering methods, such as CBC (Pantel, 2003), to identify multiple negative categories across the target categories in a single iteration. We would also like to explore the suitability of NEG-FINDER for relation extraction.

Our initial analysis demonstrated that although excellent performance is achieved using negative categories, large performance variations occur when using categories crafted by different domain experts.

In NEG-FINDER, unsupervised clustering approaches are exploited to automatically discover

	401-600	801-1000
WMEB-DRIFT	66.9	58.5
+negative 1	73.1	61.7
NEG-FINDER		
+mixture	71.9	64.2
+mixture +negative 1	76.1	66.7
WMEB-DRIFT		
+restart +mixture	78.0	70.8

Table 10: Random seed results

negative categories during bootstrapping. NEG-FINDER identifies cohesive negative categories and many of these are semantically similar to those identified by domain experts.

NEG-FINDER significantly outperforms the state-of-the-art algorithm WMEB-DRIFT, before negative categories are crafted, by up to 5.4% over the top-1000 terms; and by 10.7% over the last 200 terms extracted, where semantic drift is extensive. The new discovered categories can also be fully exploited in bootstrapping, where they successfully outperform a domain expert's negative categories and approach that of another expert.

The result is an effective approach that can be incorporated within any bootstrapper. NEG-FINDER successfully removes the necessity of including manually crafted supervised knowledge to boost a bootstrapper's performance. In doing so, we revert the multi-category bootstrapping framework back to its originally intended minimally supervised framework, with little performance trade-off.

Acknowledgements

We would like to thank Dr Cassie Thornley, our second evaluator; and the anonymous reviewers for their helpful feedback. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

Andrew Carlson, Justin Betteridge, Richard C. Wang, Jr. Estevam R. Hruschka, and Tom M. Mitchell. 2010. Coupled semi-supervised learning for information extraction. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 101–110, New York, NY, USA.

Paul R. Cohen. 1995. *Empirical Methods for Artificial Intelligence*. MIT Press, Cambridge, MA, USA.

James R. Curran, Tara Murphy, and Bernhard Scholz. 2007. Minimising semantic drift with mutual exclusion bootstrapping. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 172–180, Melbourne, Australia.

James R. Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, University of Edinburgh, Edinburgh, UK.

Claire Grover, Michael Matthews, and Richard Tobin. 2006. Tools to address the interdependence between tokenisation and standoff annotation. In *Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing*, pages 19–26, Trento, Italy.

William Hersh, Aaron M. Cohen, Lynn Ruslen, and Phoebe M. Roberts. 2007. TREC 2007 Genomics track overview. In *Proceedings of the 16th Text REtrieval Conference*, Gaithersburg, MD, USA.

Leonard Kaufmann and Peter J. Rousseeuw. 1990. *Finding Groups in Data: an Introdution to Cluster Analysis.* John Wiley and Sons.

Winston Lin, Roman Yangarber, and Ralph Grishman. 2003. Bootstrapped learning of semantic classes from positive and negative examples. In *Proceedings of the ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data*, pages 103–111, Washington, DC, USA.

Tara McIntosh and James R. Curran. 2008. Weighted mutual exclusion bootstrapping for domain independent lexicon and template acquisition. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 97–105, Hobart, Australia.

Tara McIntosh and James R. Curran. 2009. Reducing semantic drift with bagging and distributional similarity. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 396–404, Suntec, Singapore.

Tara McIntosh. 2010. *Reducing Semantic Drift in Biomedical Lexicon Bootstrapping*. Ph.D. thesis, University of Sydney.

Marius Paşca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain. 2006. Names and similarities on the web: Fact extraction in the fast lane. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 809–816, Sydney, Australia.

Patrick Pantel, Eric Crestan, Arkady Borkovsky, Ana-Maria Popescu, and Vishnu Vyas. 2009. Web-scale distributional similarity and entity set expansion. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 238–247, Singapore, Singapore.

Patrick Pantel. 2003. *Clustering by Committee*. Ph.D. thesis, University of Alberta.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 41–47, Philadelphia, PA, USA.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference*, pages 474–479, Orlando, FL, USA.

Ellen Riloff and Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124, Providence, RI, USA.

Michael Thelen and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 214–221, Philadelphia, PA, USA.

Roman Yangarber, Winston Lin, and Ralph Grishman. 2002. Unsupervised learning of generalized names. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 1135–1141, San Francisco, CA, USA.

Hong Yu and Eugene Agichtein. 2003. Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, 19(1):i340–i349.