# Relation Guided Bootstrapping of Semantic Lexicons

**Tara McIntosh**[♠] **Lars Yencken**[♠] **James R. Curran**[♢] **Timothy Baldwin**[♠]

♠ NICTA, Victoria Research Lab
Dept. of Computer Science and Software Engineering
The University of Melbourne

nlp@taramcintosh.org
lars@yencken.org

♢ School of Information Technologies
The University of Sydney

james@it.usyd.edu.au
tb@ldwin.net

## Abstract

State-of-the-art bootstrapping systems rely on expert-crafted semantic constraints such as negative categories to reduce semantic drift. Unfortunately, their use introduces a substantial amount of supervised knowledge. We present the Relation Guided Bootstrapping (RGB) algorithm, which simultaneously extracts lexicons and open relationships to guide lexicon growth and reduce semantic drift. This removes the necessity for manually crafting category and relationship constraints, and manually generating negative categories.

## 1 Introduction

Many approaches to extracting semantic lexicons extend the unsupervised bootstrapping framework (Riloff and Shepherd, 1997). These use a small set of seed examples from the target lexicon to identify contextual patterns which are then used to extract new lexicon items (Riloff and Jones, 1999).

Bootstrappers are prone to semantic drift, caused by selection of poor candidate terms or patterns (Curran et al., 2007), which can be reduced by semantically constraining the candidates. Multi-category bootstrappers, such as NOMEN (Yangarber et al., 2002) and WMEB (McIntosh and Curran, 2008), reduce semantic drift by extracting multiple categories simultaneously in competition.

The inclusion of manually-crafted negative categories to multi-category bootstrappers achieves the best results, by clarifying the boundaries between categories (Yangarber et al., 2002). For example, `female names` are often bootstrapped with the negative categories `flowers` (e.g. *Rose*, *Iris*) and `gem stones` (e.g. *Ruby*, *Pearl*) (Curran et al., 2007). Unfortunately, negative categories are difficult to design, introducing a substantial amount of human expertise into an otherwise unsupervised framework. McIntosh (2010) made some progress towards automatically learning useful negative categories during bootstrapping.

In this work we identify an unsupervised source of semantic constraints inspired by the *Coupled Pattern Learner* (CPL, Carlson et al. (2010)). In CPL, relation bootstrapping is coupled with lexicon bootstrapping in order to control semantic drift in the target relation's arguments. Semantic constraints on categories and relations are manually crafted in CPL. For example, a candidate of the relation IsCEOOf will only be extracted if its arguments can be extracted into the `ceo` and `company` lexicons and a `ceo` is constrained to not be a `celebrity` or `politician`. Negative examples such as IsCEOOf(*Sergey Brin*, *Google*) are also introduced to clarify boundary conditions. CPL employs a large number of these manually-crafted constraints to improve precision at the expense of recall (only 18 IsCEOOf instances were extracted). In our approach, we exploit open relation bootstrapping to minimise semantic drift, without any manual seeding of relations or pre-defined category lexicon combinations.

Orthogonal to these seeded and constraint-based methods is the relation-independent *Open Information Extraction* (OpenIE) paradigm. OpenIE systems, such as TextRunner (Banko et al., 2007), define neither lexicon categories nor predefined relationships. They extract relation tuples by exploit-



*Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers*, pages 266–270,
Portland, Oregon, June 19-24, 2011. 

ing broad syntactic patterns that are likely to indicate relations. This enables the extraction of interesting and unanticipated relations from text. However these patterns are often too broad, resulting in the extraction of tuples that do not represent relations at all. As a result, heavy (supervised) post-processing or use of supervised information is necessary. For example, Christensen et al. (2010) improve TEXTRUNNER precision by using deep parsing information via semantic role labelling.

## 2 Relation Guided Bootstrapping

Rather than relying on manually-crafted category and relation constraints, *Relation Guided Bootstrapping* (RGB) automatically detects, seeds and bootstraps open relations between the target categories. These relations anchor categories together, e.g. ISCEOOF and ISFOUNDEROF anchor `person` and `company`, preventing them from drifting into other categories. Relations can also identify new terms. We demonstrate that this relation guidance effectively reduces semantic drift, with performance approaching manually-crafted constraints.

RGB can be applied to any multi-category bootstrapper, and in these experiments we use WMEB (McIntosh and Curran, 2008), as shown in Figure 1. RGB alternates between two phases of WMEB, one for terms and the other for relations, with a one-off relation discovery phase in between.

### Term Extraction

The first stage of RGB follows the term extraction process of WMEB. Each category is initialised by a set of hand-picked seed terms. In each iteration, a category's terms are used to identify candidate patterns that can match the terms in the text. Semantic drift is reduced by forcing the categories to be mutually exclusive (i.e. patterns must be nominated by only one category). The remaining patterns are ranked according to *reliability* and *relevance*, and the top-$n$ patterns are then added to the pattern set.[1]

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 $\chi^2$ 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 pattern.

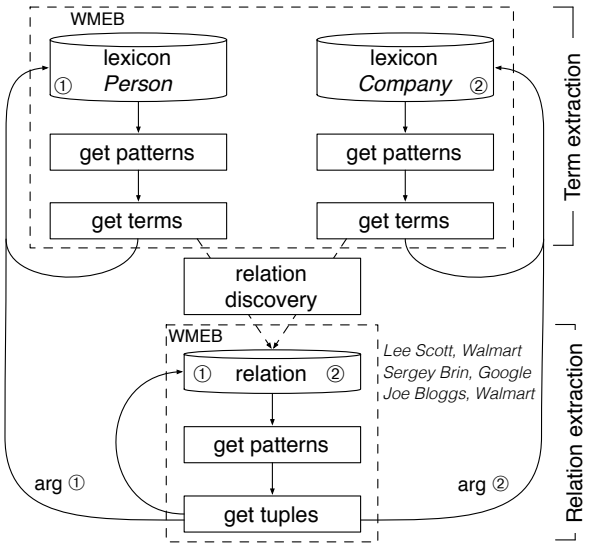


Figure 1: Relation Guided Bootstrapping framework

In WMEB's term selection phase, a category's pattern set is used to identify candidate terms. Like the candidate patterns, terms matching multiple categories are excluded. The remaining terms are ranked and the top-$n$ terms are added to the lexicon.

### Relation Discovery

In CPL (Carlson et al., 2010), a relation is instantiated with manually-crafted seed tuples and patterns. In RGB, the relations and their seeds are automatically identified in relation discovery. Relation discovery is only performed once after the first 20 iterations of term extraction, which ensures the lexicons have adequate coverage to form potential relations.

Each ordered pair of categories $(C_1, C_2) = R_{1,2}$ is checked for open (not pre-defined) relations between $C_1$ and $C_2$. This check removes all pairs of terms, *tuples* $(t_1, t_2) \in C_1 \times C_2$ with $\text{freq}(t_1, t_2) < 5$ and a cooccurrence score $\chi^2(t_1, t_2) \leq 0$.[2] If $R_{1,2}$ has fewer than 10 remaining tuples, it is discarded.

The tuples for $R_{1,2}$ are then used to find its initial set of relation patterns. Each pattern must match more than one tuple and must be mutually exclusive between the relations. If fewer than $n$ relation patterns are found for $R_{1,2}$, it is discarded. At this stage

[1] In this work, $n$ is set to 5.

[2] This cut-off is used as the $\chi^2$ statistic is sensitive to low frequencies.

| TYPE | 5gm | 5gm + 4gm | 5gm + DC |
|---|---|---|---|
| Terms | | 1 347 002 | |
| Patterns | | 4 090 412 | |
| Tuples | 2 114 243 | 3 470 206 | 14 369 673 |
| Relation Patterns | 5 523 473 | 10 317 703 | 31 867 250 |

Table 1: Statistics of three filtered MEDLINE datasets

we have identified the open relations that link categories together and their initial extraction patterns.

Using the initial relation patterns, the top-$n$ mutually exclusive seed tuples are identified for the relation $R_{1,2}$. In CPL, these tuple seeds are manually crafted. Note that $R_{1,2}$ can represent multiple relations between $C_1$ and $C_2$, which may not apply to all of the seeds, e.g. `isCeoOf` and `isEmployedBy`. We discover two types of relations, inter-category relations where $C_1 \neq C_2$, and intra-category relations where $C_1 = C_2$.

**Relation Extraction**

The relation extraction phase involves running WMEB over tuples rather than terms. If multiple relations are found, e.g. $R_{1,2}$ and $R_{2,3}$, these are bootstrapped simultaneously, competing with each other for tuples and relation patterns. Mutual exclusion constraints between the relations are also forced.

In each iteration, a relation's set of tuples is used to identify candidate relation patterns, as for term extraction. The top-$n$ non-overlapping patterns are extracted for each relation, and are used to identify the top-$n$ candidate tuples. The tuples are scored similarly to the relation patterns, and any tuple identified by multiple relations is excluded.

For tuple extraction, a relation $R_{1,2}$ is constrained to only consider candidates where either $t_1$ or $t_2$ has previously been extracted into $C_1$ or $C_2$, respectively. To extract a candidate tuple with an unknown term, the term must also be a valid candidate of its associated category. That is, the term must match at least one pattern assigned to the category and not match patterns assigned to another category.

This *type-checking* anchors relations to the categories they link together, limiting their drift into other relations. It also provides *guided term growth* in the categories they link. The growth is "guided" because the relations define, semantically coherent subregions of the category search spaces. For example, ISCEOOF defines the subregion `ceo` within `person`. This guidance reduces semantic drift.

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

Table 2: The MEDLINE semantic categories

## 3 Experimental Setup

To compare the effectiveness of RGB we consider the task of extracting biomedical semantic lexicons, building on the work of McIntosh and Curran (2008). Note however the method is equally applicable to any corpus and set of semantic categories.

The corpus consists of approximately 18.5 million MEDLINE abstracts (up to Nov 2009). The text was tokenised and POS-tagged using bio-specific NLP tools (Grover et al., 2006), and parsed using the biomedical C&C CCG parser (Rimell and Clark, 2009; Clark and Curran, 2007).

The term extraction data is formed from the raw 5-grams ($t_1$, $t_2$, $t_3$, $t_4$, $t_5$), where the set of candidate terms correspond to the middle tokens ($t_3$) and the patterns are formed from the surrounding tokens ($t_1$, $t_2$, $t_4$, $t_5$). The relation extraction data is also formed from the 5-grams. The candidate tuples correspond to the tokens ($t_1, t_5$) and the patterns are formed from the intervening tokens ($t_2, t_3, t_4$).

The second relation dataset (5gm + 4gm), also includes length 2 patterns formed from 4-grams. The final relation dataset (5gm + DC) includes dependency chains up to length 5 as the patterns between terms (Greenwood et al., 2005). These chains are formed using the Stanford dependencies generated by the Rimell and Clark (2009) parser. All candidates occurring less than 10 times were filtered. The sizes of the resulting datasets are shown in Table 1.

| | 1-500 | 501-1000 | 1-1000 |
|---|---|---|---|
| WMEB | 76.1 | 56.4 | 66.3 |
| +negative | 86.9 | 68.7 | 77.8 |
| intra-RGB | 75.7 | 62.7 | 69.2 |
| +negative | 87.4 | 72.4 | 79.9 |
| inter-RGB | 80.5 | 69.9 | 75.1 |
| +negative | 87.7 | 76.4 | 82.0 |
| mixed-RGB | 74.7 | 69.9 | 72.3 |
| +negative | 87.9 | 73.5 | 80.7 |

Table 3: Performance comparison of WMEB and RGB

| INTER-RGB | 1-500 | 501-1000 | 1-1000 |
|---|---|---|---|
| 5gm | 80.5 | 69.9 | 75.1 |
| +negative | 87.7 | 76.4 | 82.0 |
| 5gm + 4gm | 79.6 | 71.5 | 75.5 |
| +negative | 87.7 | 76.1 | 81.9 |
| 5gm + DC | 77.2 | 70.1 | 73.5 |
| +negative | 86.6 | 80.2 | 83.5 |

Table 4: Comparison of different relation pattern types

We follow McIntosh and Curran (2009) in using the 10 biomedical semantic categories and their hand-picked seeds in Table 2, and manually crafted negative categories: `amino acid`, `animal`, `body part` and `organism`. Our evaluation process involved manually judging each extracted term and we calculate the average precision of the top-1000 terms over the 10 target categories. We do not calculate recall, due to the open-ended nature of the categories.

## 4 Results and Discussion

Table 3 compares the performance of WMEB and RGB, with and without the negative categories. For RGB, we compare intra-, inter- and mixed relation types, and use the 5gm format of tuples and relation patterns. In WMEB, drift dominates in the later iterations with $\sim$19% precision drop between the first and last 500 terms. The manually-crafted negative categories give a substantial boost in precision on both the first and last 500 terms (+11.5% overall).

Over the top 1000 terms, RGB significantly outperforms the corresponding WMEB with and without negative categories ($p < 0.05$).[3] In particular, inter-RGB significantly improves upon WMEB with no negative categories (501-1000: +13.5%, 1-1000: +8.8%). In similar experiments, NEGFINDER, used during bootstrapping, was shown to increase precision by $\sim$5% (McIntosh, 2010). Inter-RGB without negatives approaches the precision of WMEB with the negatives, trailing only by 2.7% overall. This demonstrates that RGB effectively reduces the reliance on manually-crafted negative categories for lexicon bootstrapping.

The use of intra-category relations was far less effective than inter-category relations, and the combination of intra- and inter- was less effective than just using inter-category relations. In intra-RGB the categories are more susceptible to single-category drift. The additional constraints provided by anchoring two categories appear to make inter-RGB less susceptible to drift. Many intra-category relations represent listings commonly identified by conjunctions. However, these patterns are identified by multiple intra-category relations and are excluded.

Through manual inspection of inter-RGB's tuples and patterns, we identified numerous meaningful relations, such as `isExpressedIn(prot, cell)`. Relations like this helped to reduce semantic drift within the CELL lexicon by up to 23%.

Table 4 compares the effect of different relation pattern representations on the performance of inter-RGB. The 5gm+4gm data, which doubles the number of possible candidate relation patterns, performs similarly to the 5gm representation. Adding dependency chains decreased and increased precision depending on whether negative categories were used.

In Wu and Weld (2010), the performance of an OPENIE system was significantly improved by using patterns formed from dependency parses. However in our DC experiments, the earlier bootstrapping iterations were less precise than the simple 5gm+4gm and 5gm representations. Since the chains can be as short as two dependencies, some of these patterns may not be specific enough. These results demonstrate that useful open relations can be represented using only $n$-grams.

## 5 Conclusion

In this paper, we have proposed Relation Guided Bootstrapping (RGB), an unsupervised approach to discovering and seeding open relations to constrain semantic lexicon bootstrapping.

[3] Significance was tested using intensive randomisation tests.

Previous work used manually-crafted lexical and relation constraints to improve relation extraction (Carlson et al., 2010). We turn this idea on its head, by using open relation extraction to provide constraints for lexicon bootstrapping, and automatically discover the open relations and their seeds from the expanding bootstrapped lexicons.

RGB effectively reduces semantic drift delivering performance comparable to state-of-the-art systems that rely on manually-crafted negative constraints.

## Acknowledgements

We would like to thank Dr Cassie Thornley, our second evaluator, and the 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. This work has been supported by the Australian Research Council under Discovery Project DP1097291 and the Capital Markets Cooperative Research Centre.

## References

Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2670–2676, Hyderabad, India.

Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka, Jr., 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, USA.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2010. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 52–60, Los Angeles, California, USA, June.

Stephen Clark and James R. Curran. 2007. Wide-coverage efficient statistical parsing with ccg and log-linear models. *Computational Linguistics*, 33(4):493–552.

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.

Mark A. Greenwood, Mark Stevenson, Yikun Guo, Henk Harkema, and Angus Roberts. 2005. Automatically acquiring a linguistically motivated genic interaction extraction system. In *Proceedings of the 4th Learning Language in Logic Workshop*, pages 46–52, Bonn, Germany.

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.

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, August.

Tara McIntosh. 2010. Unsupervised discovery of negative categories in lexicon bootstrapping. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 356–365, Boston, 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, 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, USA.

Laura Rimell and Stephen Clark. 2009. Porting a lexicalized-grammar parser to the biomedical domain. *Journal of Biomedical Informatics*, pages 852–865.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics*, pages 118–127, Uppsala, Sweden.

Roman Yangarber, Winston Lin, and Ralph Grishman. 2002. Unsupervised learning of generalized names. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1135–1141, Taipei, Taiwan.