Combining Pattern-based and Distributional Similarity for Graph-based Noun Categorization

Michael Wiegand*, Benjamin Roth†, and Dietrich Klakow*

*Spoken Language Systems, Saarland University, D-66123 Germany
†School of Computer Science, University of Massachusetts, Amherst
michael.wiegand@lsv.uni-saarland.de

Abstract. We examine the combination of pattern-based and distributional similarity for the induction of semantic categories. Pattern-based methods are precise and sparse while distributional methods have a higher recall. Given these particular properties we use the prediction of distributional methods as a back-off to pattern-based similarity. Since our pattern-based approach is embedded into a semi-supervised graph clustering algorithm, we also examine how distributional information is best added to that classifier. Our experiments are carried out on 5 different food categorization tasks.

1 Introduction

Automatically inducing semantic categories of nouns from large unlabeled corpora is a pressing problem in natural language processing. Semantic categories are not only needed in order to build lexical ontologies, but they are also vital for relation extraction tasks in order to provide some means of generalization over traditional word-level representations.

With regard to type induction, there are two competing paradigms: *Pattern-based methods* mostly employ few hand-written surface patterns and ensure a high precision while *distributional methods* usually yield a better recall but may be considerably inferior with regard to precision.

In this paper, we examine ways to combine these methods for categorization. We apply them to 5 different tasks in the food domain (3 of which have not been addressed before) providing evidence that a combination works in general. We examine the food domain, since this domain has already been considered for natural language processing tasks [4, 5, 2, 12, 3]. Moreover, food categories have been shown to substantially improve relation extraction in this domain [23].

2 Data Set and Corpus

Since our task is to induce food categories, we need a food vocabulary as input. We use a proper subset of the food vocabulary employed in [23] where compounds

Task	Description	Categories
type	common food categories (inspired by the *Food Guide Pyramid*)	meat/fish (*pork*) 23.9, beverages (*coffee*) 13.9, spices/sauces (*cinnamon*) 12.6, sweets/pastries/snacks (*chocolate*) 12.4, vegetables/salads (*broccoli*) 9.8, starch-based side dishes (*rice*) 7.9, grains/nuts/seeds (*spelt*) 5.9, fruits (*banana*) 5.2, milk products (*cheese*) 4.2, fat (*margarine*) 2.8, eggs (*omelette*) 1.6
dish	compositionality of food items	atom (*apple*) 78.3, dish (*lasagna*) 21.7
taste	predominant taste	umami/salty (*pizza*) 56.7, sweet (*orange*) 25.8, bitter (*beer*) 6.0, sour (*vinegar*) 4.0
temperature	temperature at consumption	cold (*sandwich*) 52.2, warm (*steak*) 41.7
state of matter	state of matter at consumption	solid (*bread*) 76.5, liquid (*remoulade*) 22.5

Table 1. The categorization tasks (*each category is followed by an example and its proportion in the food vocabulary*).

have been removed.[1] It comprises 834 food items. We consider food compounds (e.g. *chocolate-almond cake*) less relevant for our investigation, since one can effectively infer (most) category labels from suffixes/heads as shown in previous work [23].[2] We want to focus on the (sparse) food items that cannot be processed with the help of this linguistic heuristic. This is a more general setting that is also relevant to other domains.

We consider the 5 different categorization tasks summarized in Table 1 addressing different properties of food items. Our food vocabulary has been annotated w.r.t. all of these categories. The first two categorization tasks have already been addressed in previous work [23], however, the remaining three tasks are examined for the first time. In each categorization task, the categories are disjoint.

Our experiments are carried out on German data. Examples are given as English translations. As an unlabeled (domain-specific) corpus from which to induce food categories, we used a crawl of *chefkoch.de* [22] consisting of 418,558 webpages of forum entries.

3 Similarity Types and Categorization

All approaches start with labeled seeds whose category labels are expanded to the remaining unlabeled items with the help of some similarity type.

3.1 Pattern-based Similarity

For pattern-based similarity, we use the *domain-independent* similarity-patterns from [23]. Each pattern is a lexical sequence that connects the mention of two

[1] We remove all food items that contain as a suffix another food item that is also contained in our food vocabulary.

[2] That is, in order to establish the label of the sparse compound *chocolate-almond cake*, one just considers the label of the suffix/head *cake*. The latter is a more general expression for which a label can be more reliably determined.

food items (Table 2). For categorization, the patterns are used to build a similarity graph, where the nodes are the food items and the edges indicate the occurrences of food items with a similarity pattern (the edge weight is the frequency of the occurrences with these patterns). Then, a semi-supervised graph clustering algorithm (as previously suggested [23]) is applied onto the graph. This requires a set of manually defined seeds for each category to be recognized. The method is a low-resource approach that only requires an unlabeled corpus and a set of seeds.

For all categorization tasks, we always employ the same similarity graph and the same graph clustering method. The only difference is the choice of seeds which represent instances of the respective categories that are to be induced.

3.2 Distributional Similarity

In order to compute distributional similarity, each food item is represented as a feature vector. The components are words that co-occur in a fixed window of 5 words (weighted by *tf-idf*) with mentions of the target food item to be represented. This vector-encoding allows all food items to be compared with each other, using the cosine-similarity. The resulting pair-wise similarities are stored in a similarity matrix (Figure 1(b)). For classification, a nearest neighbour classifier (using labeled seed food items identical to the ones from §3.1) is suitable. Such classifier has been found more effective for distributional similarity than graph-based clustering [23].

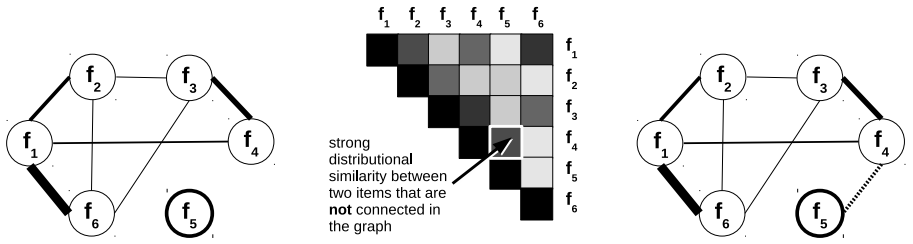
Unlike in [23], we consider k nearest neighbours rather than just the nearest neighbour. We also extend the vector representation by adding *Brown* clusters [1] of the component words to the vector representation. Brown clusters represent word clusters that are automatically induced. They have been shown to improve named-entity recognition [20] and relation extraction [15].

Patterns	food_item$_1$ (*or*|*or rather*|*instead of*|"(") food_item$_2$
Example	{*apple*: *pineapple*, *pear*, *fruit*, *strawberry*, *kiwi*} {*steak*: *schnitzel*, *sausage*, *roast*, *meat loaf*, *cutlet*}

Table 2. *Domain-independent* similarity patterns.

pattern-based similarity for		**distributional similarity for**	
asparagus (frequent term)	***kirsch (brandy)*** (rare term)	***asparagus*** (frequent term)	***kirsch (brandy)*** (rare term)
vegetable	no matching	*salsify*	*cognac*
mushroom		*salmon*	*calvados*
champignon		*chicken*	*grappa*
salsify		*pasta*	*amaretto*
salad		*savoy*	*liquor*
fish		*matjes*	*rum*

Table 3. The 6 most similar food items for two different target food items (*underlined items are unintuitive*).



(a) pattern-based graph (*line width of edge indicates similarity*)

(b) distributional similarity matrix (*darkness of cell indicates similarity*)

(c) augmented graph (*line width of edge indicates similarity strength*)

Fig. 1. Combination of pattern-based and distributional similarity (f_i *represents some food item*).

3.3 Comparing the Two Similarity Types

Pattern-based and distributional methods have complementary properties. This is illustrated by Table 3 which shows the 6 most similar food items to *asparagus* and *kirsch* according to each of the similarity types. *Asparagus* is a frequent food item (31,355 mentions in our corpus) while *kirsch* is rare (34 mentions). As a consequence, none of the similarity patterns are observed with the rare item, hence *kirsch* is an unconnected node in the graph. For unconnected nodes, graph-based clustering is unable to make a prediction. This concerns 15.8% of the food items in our vocabulary. With distributional similarity, however, we obtain similar food items for *all* food items. But Table 3 also illustrates that the *quality* (precision) of pattern-based similarity is superior to distributional similarity. This is because the similarity patterns are based on *coordination* which is known to ensure semantic coherence [25]. We, therefore, assume that distributional similarity is only helpful when pattern-based similarity provides no prediction.

3.4 Combination Methods

We examine 3 methods to combine distributional and pattern-based similarity. They all use distributional similarity as a back-off to pattern-based similarity. This should primarily mitigate the sparsity in the pattern-based graph caused by food items that are not connected to any other food item (f_5 in Figure 1(a)). For those food items, some similarity information is obtained by distributional similarity ($edge(f_4, f_5)$ in Figure 1(b)) and can, for example, be included in the similarity graph (Figure 1(c)):

- **cascade:** We run graph clustering (on the original pattern-based similarity graph) and the nearest neighbour classifier (using distributional similarity) in parallel; per default the prediction of graph clustering is taken, only if no prediction could be produced by that method, the prediction of the nearest neighbour classifier is used.
- **graph-aug**$_{local}$**:** Information from the distributional similarity matrix is directly included in the (pattern-based) graph; for each unconnected food item, edges to the n most similar food items according to the distributional similarity matrix are added.

– **graph-aug$_{global}$:** Similar to *graph-aug$_{local}$* but for *every* food item in the food vocabulary, the n most distributionally similar food items are connected by additional edges.

The first method is a naive combination that also keeps pattern-based and distributional similarity separated from each other during training, while the other two methods are integrated solutions. The purpose of the third method is to check whether even beyond food items in the graph that are not connected, additional back-off edges from distributional similarity may help. For both integrated solutions, we employ the distributional similarity score *ds* as an edge weight in the graph. *ds* is always in the range $[0; 1[$. It is therefore always smaller than the pattern-based similarity score of observed patterns *ps* (which denotes the absolute frequency of pattern occurrences), i.e. $ps > ds$ since $ps >= 1$. This encoding should reflect that we consider distributional similarity as a back-off.

4 Experiments

	without Brown				**with Brown**			
k	**1**	**3**	**5**	**10**	**1**	**3**	**5**	**10**
Acc	64.9	62.1	61.5	57.8	**67.5**	64.4	64.4	61.3
F	62.3	59.7	58.4	54.2	**64.5**	60.9	60.3	56.9

Table 4. Varying k in nearest neighbour classification and examining the impact of *Brown* cluster features (*results averaged over tasks*).

	cascade		**graph-aug$_{local}$**		**graph-aug$_{global}$**	
Task	**Acc**	**F**	**Acc**	**F**	**Acc**	**F**
type	78.66	76.96	78.78	76.22	**80.46**	**78.42**
dish	71.34	66.06	**76.74**	69.89	**76.74**	**70.96**
taste	71.47	60.95	73.15	62.73	**74.32**	**63.64**
temperature	77.14	77.07	**78.03**	**78.17**	76.88	76.80
state of matter	81.72	78.32	84.02	80.61	**84.62**	**81.19**
average	76.07	71.87	78.14	73.52	**78.60**	**74.20**

Table 5. Comparison of combining pattern-based and distributional similarity.

edges	**1**	**2**	**3**	**5**	**10**	**20**
Acc	78.09	**78.60**	78.10	77.74	77.58	75.37
F	74.15	**74.20**	73.81	73.28	73.52	71.85

Table 6. Varying the number of edges to be added in *graph-aug$_{global}$* (*results averaged over tasks*).

As seeds we randomly sampled for every category of every task (Table 1) 20 seeds. For graph-based clustering, we use the configuration of hyper-parameters from previous work [23]. We induced 1000 Brown clusters from our domain-specific corpus with SRILM [19].

Table 4 shows different configurations for nearest neighbour classification using distributional similarity. Increasing the number of nearest neighbours notably decreases performance. However, using Brown clusters as features is beneficial. Therefore, for all further experiments using a k nearest neighbour classifier, we will always set $k = 1$, however, we include Brown clusters as context features.

Table 5 compares the different methods combining pattern-based and distributional similarity. On average, the naive combination method (i.e. *cascade*)

	majority classifier				nearest neighbour *(distributional similarity)*				graph *(pattern-based similarity)*				graph-aug$_{global}$ *(combination)*			
Task	**Acc**	**Prec**	**Rec**	**F**	**Acc**	**Prec**	**Rec**	**F**	**Acc**	**Prec**	**Rec**	**F**	**Acc**	**Prec**	**Rec**	**F**
type	23.9	2.2	9.1	3.5	63.4	64.0	72.2	65.4	74.7	**81.7**	79.9	**79.3**	**80.5**	75.4	**84.3**	78.4
dish	**78.3**	39.2	50.0	43.9	64.2	60.5	65.1	59.1	63.2	68.4	63.9	63.8	76.7	**69.6**	**75.8**	**71.0**
taste	61.4	15.3	25.0	19.0	57.1	49.5	66.8	49.7	64.2	**62.0**	69.9	61.4	**74.3**	59.7	**76.8**	**63.6**
temperature	55.6	27.8	50.0	35.7	75.0	75.0	74.0	74.2	67.0	**79.6**	67.4	72.7	**76.9**	76.9	**77.2**	**76.8**
stater of mat.	77.2	38.6	50.0	43.6	78.0	72.8	80.7	74.0	72.6	**81.0**	75.9	76.6	**84.6**	79.0	**87.2**	**81.2**
average	59.3	24.6	36.8	29.1	67.5	64.4	71.7	64.5	68.4	**74.6**	71.4	70.8	**78.6**	72.1	**80.3**	**74.2**

Table 7. Comparison of different methods.

performs worst. The best overall result is obtained by the integrated solution with the global edge extension (i.e. *graph-aug*$_{global}$).

For the integrated methods in Table 5, we always used the 2 most similar items from the distributional similarity matrix. Table 6 shows that for this value we obtained maximum performance.

Table 7 compares the best combination method against the original graph clustering, nearest neighbour and majority-class classifier. For most tasks, the combination outperforms the best individual classifier (*nearest neighbour*/*graph*).

The improvement in F-score by combining pattern-based and distributional similarity is most notably caused by raising recall. The combined approach largely outperforms the majority-class classifier w.r.t. F-score. (In terms of accuracy, there is only one task, i.e. *dish*, in which that baseline is not beaten.) The proposed method also produces reasonable results on the new categorization tasks not previously examined (i.e. *taste*, *temperature* and *state of matter*).

5 Related Work

The types of categorizations we present in this paper are typical instances of noun classification. For that task, both distributional methods [16, 11, 18, 21, 24, 7, 10] and pattern-based methods [6, 14, 9, 8] have been explored. The complementarity of those methods has only been examined for textual entailment [13] and categorization of *raw semantic classes* [17]. While our paper is the first work that combines these methods in the context of graph-based clustering, those previous publications consider different classification methods, i.e. supervised learning and query set expansion, that require a different combination.

This work also extends the types of categorizations applied on the food domain addressing *taste*, *state of matter* and *temperature* for the first time.

6 Conclusion

We presented a combined approach for the induction of noun categories using pattern-based and distributional similarity. We considered various food categorization tasks, including three novel tasks. The best combination is a clustering approach on a pattern-based graph that also includes for each food item edges to the two most similar food items according to distributional similarity. This method outperforms both mere pattern-based and distributional methods.

Acknowledgements

This work was supported, in part, by the German Federal Ministry of Education and Research (BMBF) under grant no. 01IC12SO1X and the Information Extraction and Synthesis Lab at the University of Massachusetts. The authors would like to thank Stephanie Köser for annotating the dataset presented in this paper.

References

1. Brown, P.F., deSouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Computational Linguistics 18, 467–479 (1992)
2. Chahuneau, V., Gimpel, K., Routledge, B.R., Scherlis, L., Smith, N.A.: Word Salad: Relating Food Prices and Descriptions. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL). pp. 1357–1367. Jeju Island, Korea (2012)
3. Druck, G., Pang, B.: Spice it up? Mining Refinements to Online Instructions from User Generated Content. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 545–553. Jeju, Republic of Korea (2012)
4. van Hage, W.R., Katrenko, S., Schreiber, G.: A Method to Combine Linguistic Ontology-Mapping Techniques. In: Proceedings of International Semantic Web Conference (ISWC). pp. 732 – 744. Springer, Galway, Ireland (2005)
5. van Hage, W.R., Kolb, H., Schreiber, G.: A Method for Learning Part-Whole Relations. In: Proceedings of International Semantic Web Conference (ISWC). pp. 723 – 735. Springer, Athens, GA, USA (2006)
6. Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: Proceedings of the International Conference on Computational Linguistics (COLING). pp. 539–545. Nantes, France (1992)
7. Huang, R., Riloff, E.: Inducing Domain-specific Semantic Class Taggers from (almost) Nothing. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 275–285. Uppsala, Sweden (2010)
8. Kozareva, Z., Hovy, E.: Semi-Supervised Method to Learn and Construct Taxonomies using the Web. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1110–1118. Cambridge, MA , USA (2010)
9. Kozareva, Z., Riloff, E., Hovy, E.: Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 1048–1056. Columbus, OH, USA (2008)
10. Lenci, A., Benotto, G.: Identifying hypernyms in distributional semantic spaces. In: Proceedings of the Joint Conference on Lexical and Computational Semantics (*SEM). pp. 75–79. Montréal, Quebec, Canada (2012)
11. Lin, D.: Automatic Retrieval and Clustering of Similar Words. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics (ACL/COLING). pp. 768–774. Montreal, Quebec, Canada (1998)
12. Miao, Q., Zhang, S., Zhang, B., Meng, Y., Yu, H.: Extracting and Visualizing Semantic Relationships from Chinese Biomedical Text. In: Proceedings of the Pacific Asia Conference on Language, Information and Compuation (PACLIC). pp. 99–107. Bali, Indonesia (2012)

13. Mirkin, S., Dagan, I., Geffet, M.: Integrating Pattern-based and Distributional Similarity Methods for Lexical Entailment Acquisition. In: Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL). pp. 579–586. Sydney, Australia (2006)
14. Pantel, P., Ravichandran, D., Hovy, E.: Towards Terascale Knowledge Acquisition. In: Proceedings of the International Conference on Computational Linguistics (COLING). pp. 771–777. Geneva, Switzerland (2004)
15. Plank, B., Moschitti, A.: Embedding Semantic Similarity in Tree Kernels for Domain Adapation of Relation Extraction. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 1498–1507. Sofia, Bulgaria (2013)
16. Riloff, E., Shepherd, J.: A Corpus-Based Approach for Building Semantic Lexicons. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 117–124. Providence, RI, USA (1997)
17. Shi, S., Zhang, H., Yuan, X., Wen, J.R.: Corpus-based Semantic Class Mining: Distributional vs. Pattern-Based Approaches. In: Proceedings of the International Conference on Computational Linguistics (COLING). pp. 993–1001. Beijing, China (2010)
18. Snow, R., Jurafsky, D., Ng, A.Y.: Learning Syntactic Patterns for Automatic Hypernym Discovery. In: Advances in Neural Information Processing Systems (NIPS). Vancouver, British Columbia, Canada (2004)
19. Stolcke, A.: SRILM - An Extensible Language Modeling Toolkit. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP). pp. 901–904. Denver, CO, USA (2002)
20. Turian, J., Ratinov, L., Bengio, Y.: Word Representations: A Simple and General Method for Semi-supervised Learning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 384–394. Uppsala, Sweden (2010)
21. Weeds, J., Weir, D., McCarthy, D.: Characterising Measures of Lexical Distributional Similarity. In: Proceedings of the International Conference on Computational Linguistics (COLING). pp. 1015–1021. Geneva, Switzerland (2004)
22. Wiegand, M., Roth, B., Klakow, D.: Web-based Relation Extraction for the Food Domain. In: Proceedings of the International Conference on Applications of Natural Language Processing to Information Systems (NLDB). pp. 222–227. Springer, Groningen, the Netherlands (2012)
23. Wiegand, M., Roth, B., Klakow, D.: Automatic Food Categorization from Large Unlabeled Corpora and Its Impact on Relation Extraction. In: Proceedings of the Conference on European Chapter of the Association for Computational Linguistics (EACL). pp. 673–682. Gothenburg, Sweden (2014)
24. Yamada, I., Torisawa, K., Kazama, J., Kuroda, K., Murata, M., Saeger, S.D., Bond, F., Sumida, A.: Hypernym Discovery Based on Distributional Similarity and Hierarchical Structures. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 929–927. Singapore (2009)
25. Ziering, P., van der Plas, L., Schuetze, H.: Bootstrapping Semantic Lexicons for Technical Domains. In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP). Nagoya, Japan (1321–1329 2013)