Web of Wine Words: Hierarchy Visualization of Wine Speak by Restricted Bootstrap

Brendan Flanagan[1], **Sachio Hirokawa**[2]
[1]Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan
[2]Research Institute for Information Technology, Kyushu University, Fukuoka, Japan

Abstract - *Visualization of the relation of characteristic words can be useful for interpreting search results and enable comparisons to be made between multiple searches. In this paper we introduce a method of analysis by applying restricted bootstrapping to a set of characteristic words for extracting specificity or generality relations. These relations are used to construct a tree structure of the characteristic words that represents their hierarchical specificity or generality. This method was applied to a corpus of wine tasting notes to identify the characteristics of two wine regions by hierarchy tree. The results are compared with the frequencies of the characteristic words.*

Keywords: Restricted Bootstrap, Characteristic Hierarchy Visualization, Topic Drift, Wine Speak

1 Introduction

When searching it can be difficult to interpret from the results whether a characteristic has specificity or generality to the query. This is particularly apparent when trying to compare characteristics of search results for queries that might share many common attributes, such as wines from different regions. The characteristics of wine are often described using specialist expressions, called wine speak. People not familiar with these types of descriptions can find them confusing, as they don't have an understanding of different levels of particular and general expressions. The relations of characteristics are not immediately obvious when comparing simple search results. Some characteristics could be more of general quality, whereas others might be specific to the search query. The characteristics of search results can be thought of as a hierarchy tree, with words that are similar attaching to the same parent node. The parent nodes are then connect to create a word tree in order of greatest generality to the query at the root and greatest similarity at the lowest leaf nodes.

In this paper, we will demonstrate how the restricted bootstrap algorithm can be used to extract the degree of common generality between a pair of words. We investigate a method of generating a hierarchy of words from the specificity and generality relation of characteristic words and a corpus of target documents.

2 Related work

2.1 Information Extraction by Bootstrap

In previous research, Mihalcea et al. [7] and Palshikar [8] analyzed networks of words as undirected graphs made from the results of query word searches. Pantel et al. [9] proposed a minimally supervised bootstrapping algorithm named Espresso to extract semantic relations. Komachi et al. [6] demonstrated that semantic drift in bootstrapping is similar to that of the HITS algorithm and proposed graph based methods based on von Neumann kernels and regularized Laplacian to reduce semantic drift. While the above researches were carried out independently of each other, Radev et al. [12] notes that they all are based on the analysis of word co-occurrence as a bi-partite graph of documents and words by traversing back and forth between different node types to find feature words and sentences. We have previously revealed in [2] that semantic drift can be reduced by restricting the query length at each iteration of the bootstrap process. This algorithm was used to visualize the generality relation of search results in [4]. Also in other research, [3] we have examined the relations of wine speak expressions found in wine magazine blogs and visualized these as mind maps to support the learning of wine speak. In the present paper, we propose a method of extracting the specificity or generality relation of characteristic words to a search query by applying restricted bootstrapping.

2.2 Wine Tasting Note Analysis

There are many papers on research into the language that is used to describe wines, called Wine Speak. Some of this research is dedicated to analyzing wine tasting notes from different points of view. Caballero [1] focused on how manner-of-motion verbs are used from the point of view of describing a wine's intensity and persistence. Manner-of-motion verbs occur often used in wine tasting notes to depict motions, such as "hints of milk chocolate and vanilla *sneak in* on the palate". A corpus of wine tasting collected from the Wine Enthusiast, Wine Spectator, and Wine Advocate was analyzed and examples of 56 typical sentences that contain such verbs were given. Paradis [10] investigated the analysis of semantic middles in wine tasting notes and their use as a recommender to estimate prime drinking time. A sub corpus of 200 notes was randomly selected from a corpus of 80,000 notes from the Wine Advocate and a meticulous evaluation of

38 sentences was given. There is also related research into the visualization of wine tasting notes for linguistic analysis. Kerren et al. [5] visualized wine tasting notes using word trees generated from parts of speech and words. Their system enables the analysis of linguistic patterns within single wine reviews or based on regions and varieties. The system is highly specialized with the intention to be used for linguistic exploration of wine tasting notes. In previous research, we examined the relations of wine speak expressions found in wine magazine blogs and visualized these as mind maps to support the learning of wine speak [4]. In the present paper, we propose a method of using the restricted bootstrapping algorithm to search for common generality between pairs of query words. This is then analyzed to generate a word hierarchy of the query words with relation to the target documents in the corpus.

3 Hierarchy generation by restricted bootstrapping

The generation of characteristic word hierarchies by restricted bootstrapping involves two main steps: firstly, applying restricted bootstrapping for each characteristic word paired with the target query, and secondly, generating the characteristic word hierarchy based on the analysis of the restricted bootstrapping results.

3.1 Restricted Bootstrap Algorithm

The second author of this paper initially proposed a method of restricted bootstrapping [2] as a solution to the problem of semantic drift that sometimes occurs when the result of a bootstrap that has been applied to documents and words is too far from the initial query. When evaluated on the extraction of infrequent characteristic words, we confirmed that a tight bootstrap restriction of k = 1 produces a 10% increase in the Mean Average Precision when compared to a looser restriction of k = 50. However a comprehensive quantitative evaluation is still required as future work.

```
BS(U,q,k) {
 W = {}
 i=0
 while(true){
    Wi = word(doc(U && q))
    W = top(k,Wi)
    last if W == U
    i = i+1
    U = W
 }
 return W
}
```

Figure 1. Restricted Bootstrap Algorithm

In this paper, the restricted bootstrap algorithm *BS(U,q,k)*, shown in Fig. 1, extracts words for each word in the characteristic word set U with relation to the initial query q. The algorithm starts by searching for a query comprising of the initial query q and a word from the characteristic word set U. The words of the documents in the search results are then ranked. As the algorithm was realized using a search engine constructed using GETA[1], the default SMART weight [13] was used as the word score for ranking each word. The bootstrap length restriction k is used to limit the number of top ranking search result words that are then used as the query for the next iteration of the bootstrap process. The iteration continues until the bootstrap has converged, which is when the current k top ranked words are the same as the k top ranked words from a previous iteration. The top k ranked words of the last iteration are returned as the results of the restricted bootstrap process.

3.2 Hierarchy Tree Generation

The bootstrap results of words that match at small k are closely related to the initial query. Conversely, a larger k increases the possibility of topic drift. We propose that these properties of restricted bootstrap can be used to analyze the relations of words within a corpus.

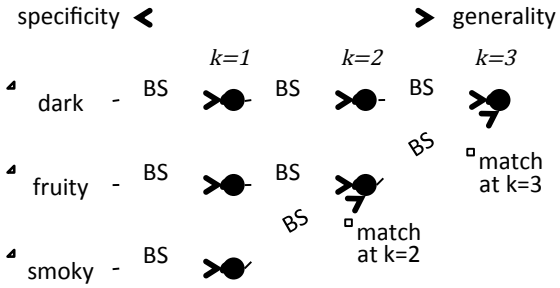

Figure 2. Hierarchy Tree Generation

An overview of the hierarchy tree generation process is shown in Fig. 2. The bootstrap results for all characteristic words are checked for exact matches at each step of increasingly larger k restriction lengths. Exact matches with the smallest restriction length k represent the relation between two or more characteristic words and are linked to the same parent node that represents the k search step. Small k represents a strong relation and large k a weak relation to the initial search query. Small k nodes are linked with the next largest k node, until only one last k node is reached. We constructed a directed graph G = (N,E) of 17 characteristic words U of wine speak given a query q as follows where N is the set of nodes as defined in Equation 1, and E is the set of edges as defined in Equation 2.

$$N = \{BS(\{w_i\}, q, k) | i = 1,..,17; k = 0,1,\dots;\} \quad (1)$$

$$\begin{aligned} \mathrm{E} = \{&BS(\{u\}, q, m), BS(\{v\}, q, n) | \\ &\exists l \geq m \; s.t. \; BS(\{u\}, q, l) = BS(\{v\}, q, n) \\ &BS(\{v\}, q, n) = BS(\{u\}, q, l_0) \\ &l_0 = \min\{l | BS(\{v\}, q, n)\}\} \qquad \} \end{aligned} \quad (2)$$

[1] http://geta.ex.nii.ac.jp/

This method can essentially be thought of as drawing paths of restricted bootstrapping results of increasingly larger k for each characteristic word in the set U. The paths are then merged at the point of an exact match for the lowest existing bootstrap restriction k.

4 Examples of hierarchy trees generated by applying restricted bootstrapping

4.1 Data Collection

A prototype system of the method was applied to a corpus consisting of 91,010 wine tasting notes from the Wine Enthusiast Magazine's Buying Guide[2] website. The attributes of each wine and the tasting notes were collected. We constructed a search engine to analyze the wine tasting notes, with each note containing an average of 2.8 sentences made up of 40 words. As an example of the method proposed in this paper, wine attributes of two regions: New Zealand and Marlborough were selected as the focus for analysis.

4.2 Characteristic Words: Sensory Expression

A list of 17 sensory modalities grouped in three categories: vision (purple, ruby, straw, gold, light, dark), smell (fruity, floral, spicy, smoky, weak), taste & touch (flabby, soft, heavy, thin, long, crisp), from Paradis and Eeg-Olofsson [11] were chosen as the set of characteristic words U. The authors have previously used these words in the analysis of wine blogs [3].

4.3 Comparison of Wine Regions

A naïve analysis method would be to compare the frequency distributions of the characteristic word set, which is shown in Fig. 3 and Fig. 4. Words of a high frequency would have a stronger relation to the corpus than words of lower frequency. This is a simple ranking method when compared to the restricted bootstrap method proposed in this paper. This is because it doesn't take into account whether the characteristic keyword U are specific or general with relation to the query q.

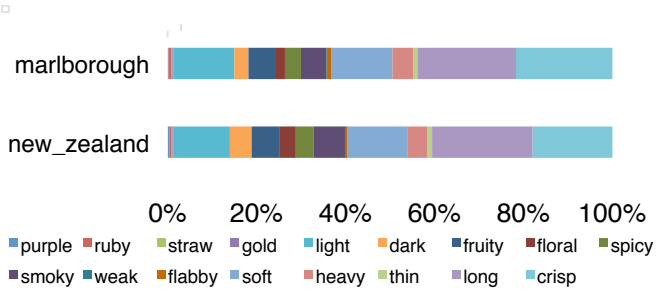

Figure 3. Comparison of region by word frequency

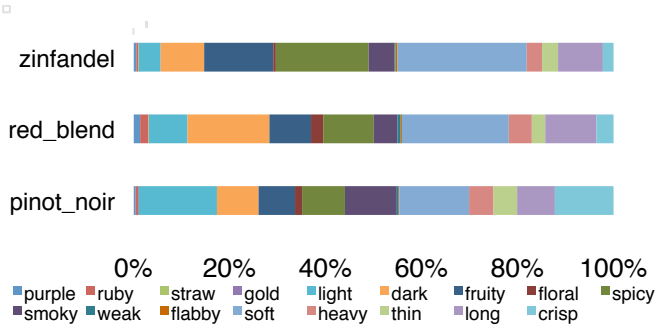

Figure 4. Comparison of grape variety by word frequency

A matching restricted bootstrapping search was applied with k from length 1 to 100 for each characteristic word to the tasting notes for wines from New Zealand and Marlborough (which is a famous region within New Zealand) containing 1427 and 715 documents respectively. The bootstrap results of all the possible pairs of characteristic words matched at least once before $k = 100$.

In Fig. 5 and 6 the hierarchy word trees produced with the proposed method for wine tasting notes from q = Marlborough and q = New Zealand are shown respectively, where the words in the nodes are colored red, green and blue for each category. The hierarchy tree is drawn from left to right as the bootstrap restriction length k increases, with words on the left side of the graph having a stronger relation to the corpus than words on the right side of the graph.

The five least frequent characteristic words U in the New Zealand corpus in ascending frequency order are: *straw*, *weak*, *gold*, *ruby*, and *flabby*. However these words are in the middle of the hierarchy tree suggesting that they have a stronger characteristic relation to New Zealand wines than would be expected by looking at the word distribution. The same can also be seen in the six least frequent characteristic words U for Marlborough in ascending order are: *purple*, *straw*, *gold*, *weak*, *dark*, and *flabby*. These words also occur in the middle of the hierarchy tree suggesting a stronger relation than could be deduced from the frequency distributions.

The distribution for the characteristic word *floral* has no difference between New Zealand and Marlborough by sample size rank (both rank 9), where $rank(w_i) = \#\{w_j \in W | w_i > w_j\}$ for the set of characteristic words W. However, in the hierarchy word tree, the difference between the depths of the *floral* node from the root of the trees is of 5 nodes. The *floral* node is closer to the root of the tree for the Marlborough corpus, which indicates that the minimum k matching restricted bootstrap results occurred at a larger k than found in the New Zealand corpus. This indicates that *floral* is a stronger characteristic of New Zealand wines than those from Marlborough, which is not apparent when comparing the distributions of words.

[2] http://buyingguide.winemag.com/

5 Conclusion and future work

In this paper, we proposed that by applying restricted bootstrapping for a set of characteristic words to a corpus of documents, a hierarchy tree representing the specificity and generality relation of the characteristics could be generated. This method was then applied to a corpus of wine tasting notes as an example of the analysis of differences in sensory expression characteristics of wine regions. Hierarchy trees were generated using the proposed matching restricted bootstrap search method for two wine regions. The results were then compared to the frequencies of the characteristic words as a naïve analysis baseline.

In future work, we plan to investigate the influence that the corpus size has on the relations of characteristics and the generated hierarchy trees. A formal method is also required to evaluate the effectiveness of extracting and generating specificity and generality relations of characteristic word sets.

6 Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 24500176.

7 References

[1] Caballero, R. 2007. Manner-of-motion verbs in wine description. *Journal of Pragmatics*, 39, 12, 2095-2114.

[2] Hirokawa, S. 2012. Feature Extraction Using Restricted Bootstrapping. *ICIS2013*, 283-288.

[3] Hirokawa, S., Flanagan, B., Suzuki, T., Yin, C. 2014. Learning Winespeak from Mind Map of Wine Blogs. In S. Yamamoto (Ed.): *Human Interface and the Management of Information Part II* (Springer LNCS 8522), 383-393.

[4] Hirokawa, S., Flanagan, B., Yin, C., Nakae, H. 2014. Visualization of relation and generality of words in research results. *ACIS2014*, 90-95.

[5] Kerren, A., Prangova, M., Paradis, C. 2011. Visualization of sensory perception descriptions. *Proc. of the 2011 15th International Conference on Information Visualisation IEEE*, 135-144.

[6] Komachi, M., Kudo, T., Shimbo, M., Matsumoto, Y. 2008. Graph-based analysis of semantic drift in Espresso-like bootstrapping algorithms. In *Proc. of the Conference on Empirical Methods in Natural Language Processing* (EMNLP 2008), 1011-1020.

[7] Mihalcea, R., & Tarau, P. 2004. TextRank: Bringing order into texts. *Proc. of Conference on Empirical Methods in Natural Language Processing* (EMNLP'2004), 404–411.

[8] Palshikar, G. K. 2007. Keyword extraction from a single document using centrality measures. *In Pattern Recognition and Machine Intelligence*, Springer LNCS 4815, 503-510.

[9] Pantel, P., & Pennacchiotti, M. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proc. of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 113-120.

[10] Paradis, C. 2009. "This beauty should drink well for 10–12 years": a note on recommendations as semantic middles. Text and Talk - An Interdisciplinary Journal of Language, *Discourse Communication Studies*, 29, 1, 53-73.

[11] Paradis, C., Eeg-Olofsson, M. 2013. Describing Sensory Experience: The Genre of Wine Reviews. *Metaphor and Symbol*, 28, 1, 22-40.

[12] Radev, D. R., Mihalcea, R. 2008. Networks and natural language processing. *AI magazine*, 29, 3, 16-28.

[13] Salton, G., McGill, M. J. 1983. *Introduction to modern information retrieval*. McGraw-Hill.

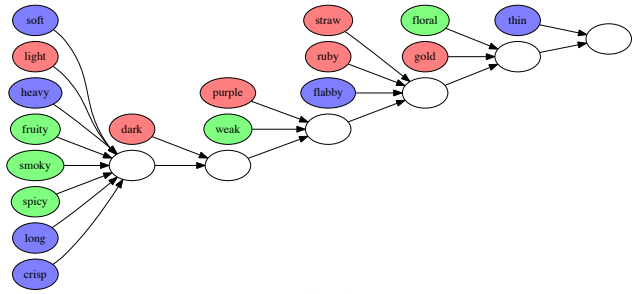

Figure 5. Hierarchy word tree of query word q = Marlborough

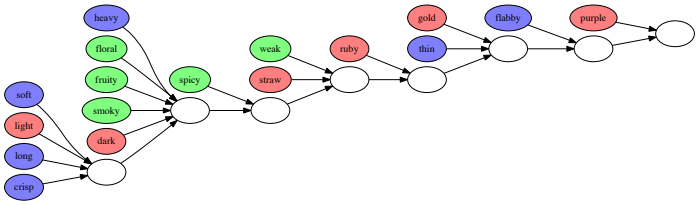

Figure 6. Hierarchy word tree of query word q = New Zealand