Generalized Representation of Syntactic Structures

Reihane Boghrati (boghrati@usc.edu)
Department of Computer Science, University of Southern California
Los Angeles, CA 90089 USA

Kate M. Johnson (katejohn@usc.edu)
Department of Psychology, University of Southern California
Los Angeles, CA 90089 USA

Morteza Dehghani (mdehghan@usc.edu)
Department of Psychology and Department of Computer Science, University of Southern California
Los Angeles, CA 90089 USA

Abstract

Analysis of language provides important insights into the underlying psychological properties of individuals and groups. While the majority of language analysis work in psychology has focused on semantics, psychological information is encoded not just in what people say, but how they say it. In the current work, we propose Conversation Level Syntax Similarity Metric-Group Representations (CASSIM-GR). This tool builds generalized representations of syntactic structures of documents, thus allowing researchers to distinguish between people and groups based on syntactic differences. CASSIM-GR builds off of Conversation Level Syntax Similarity Metric by applying spectral clustering to syntactic similarity matrices and calculating the center of each cluster of documents. This resulting cluster centroid then represents the syntactical structure of the group of documents. To examine the effectiveness of CASSIM-GR, we conduct three experiments across three unique corpora. In each experiment, we calculate the clustering accuracy and compare our proposed technique to a bag-of-words approach. Our results provide evidence for the effectiveness of CASSIM-GR and demonstrate that combining syntactic similarity and tf-idf semantic information improves the total accuracy of group classification.

Keywords: Syntax; Text Clustering; Syntactic Similarity; Text Classification; CASSIM.

Introduction

Language lies at the heart of human communication, and analysis of language has been shown to be an essential lens for investigating and understanding many different psychological properties. Language analysis has provided insight into depression (Ramirez-Esparza, Chung, Kacewicz, & Pennebaker, 2008), moral values (Graham, Haidt, & Nosek, 2009; Dehghani et al., 2016), neuroticism and extraversion (Mehl, Robbins, & Holleran, 2012), political orientations (Dehghani, Sagae, Sachdeva, & Gratch, 2014), and cultural backgrounds (Maass, Karasawa, Politi, & Suga, 2006; Dehghani, Bang, et al., 2013) among many others.

Most of these studies, however, focus on quantifying word choice or semantics. While semantics undoubtedly play an important role in capturing psychological properties, it is vital to also include analysis of syntax in this process. Prior research has shown that syntactic structures also capture individuals and group differences for various demographic and psychological factors such as educational or regional background (Bresnan & Hay, 2008), gender (Vigliocco & Franck, 1999), socio-economics (Jahr, 1992), and emotional states and personality (Gawda, 2010).

Recently, several tools have been developed for automated analysis of syntactic structures. For example, Lu's (Lu, 2010) system analyzes fourteen different measures including the ratio of verb phrases, number of dependent clauses, and T-units to calculate documents' syntactic complexity. Similarly, TAALES relies on several features such as frequency, range, academic language, and psycholinguistic word information to measure lexical sophistication (Kyle & Crossley, 2015). By comparison, Coh-Metrix is a tool which provides measurement for over 200 different facets of syntax (e.g. mean number of modifiers per noun phrase, mean number of high-level constituents per word, and the incidence of word classes that signal logical or analytical difficulty) (Graesser, McNamara, Louwerse, & Cai, 2004).

While each of these tools provides different mechanisms for measuring various syntactic features, they all rely on previously identified features of interest. More recently, we introduced ConversAtion Level Syntax Similarity Metric (CASSIM) to incorporate constituency parse trees when calculating the syntactic similarity of documents (Boghrati, Hoover, Johnson, Garten, & Dehghani, 2017). CASSIM compares groups of documents based on underlying syntactic differences between groups of documents.

There are some situations, however, where hypothesis testing about predefined features or groups may not be the only aim. Instead, researchers may wish to identify new groupings of documents and the features which tie them together. These group-level linguistic representations can lead to important, novel discoveries about how a group communicates. Clustering techniques are widely used for this type of analysis. There is an extensive literature studying various text clustering approaches and their applications (Song, Li, & Park, 2009; Sasaki & Shinnou, 2005; Lin, Jiang, & Lee, 2014). This literature demonstrates that many linguistic features facilitate improvements in text clustering (T. Liu, Liu, Chen, & Ma, 2003; L. Liu, Kang, Yu, & Wang, 2005), some of which address the effect of synonymy, hypernymy, syntax, and part of speech tags on text clustering methods (Sedding & Kazakov, 2004; Lewis & Croft, 1989; Lewis, 1992; Zheng, Kang,

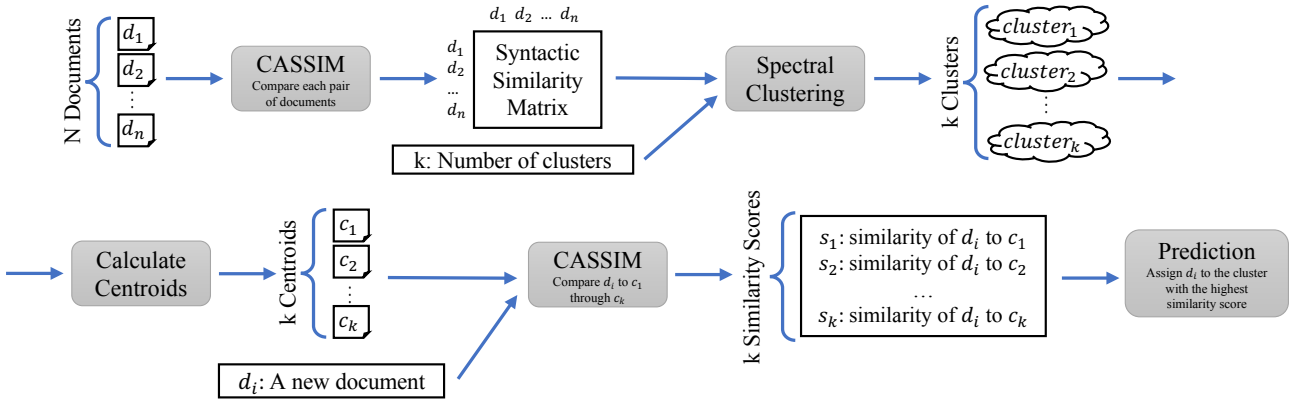

Figure 1: CASSIM-Group Representation Process.

& Kim, 2009).

In the current paper, we introduce ConversAtion Level Syntax Similarity Metric-Group Representations (CASSIM-GR), a tool that captures the generalized representation of syntactic structure used by individuals in a certain group. CASSIM-GR groups documents into separate clusters based on their syntactic similarity scores, and uses the centroid of a cluster as a generalized representation of the syntactic structures used in that cluster. These centroid syntax representation can then be used to understand within-group syntax similarities and between-group syntax variations. As we will show, these generalizations of syntactic structures can be useful when analyzing differences between documents written by different individuals or groups.

This paper is structured as follows: First, we describe our proposed approach, CASSIM-GR, in more detail. Next, we validate the approach with a corpus of syntactically similar documents. Then, we apply CASSIM-GR to two other corpora: documents marked as dogmatic and non-dogmatic (Fast & Horvitz, 2016) and documents from conservative and liberal weblogs (Dehghani, Sagae, Sachdeva, & Gratch, 2013) and evaluate the classification accuracy of CASSIM-GR compared to tf-idf approach and a combination of the two approaches. Finally, we discuss limitation and future directions of our work.

CASSIM-GR

In this section we describe CASSIM-GR for clustering groups of documents with similar syntactic structures. CASSIM-GR includes four general steps: 1. constructing the syntactic similarity matrix, 2. applying spectral clustering, 3. calculating the center of clusters, 4. classification. Figure 1 demonstrates the steps involved in CASSIM-GR to compute the generalized representation of syntactic structures.

First, we use CASSIM (Boghrati et al., 2017) to calculate the syntactic similarity between each pair of documents. CASSIM relies on edit distance difference of constituency parse trees. It first generates parse trees for the sentences in each document. Next, it calculates the edit distance between each two sentences' constituency parse trees and matches the most syntactically similar sentences using Hungarian algorithm. Finally, it provides a score between 0 and 1 where higher numbers indicate higher similarity between two documents. Using the syntactic similarity scores measured by CASSIM, we build a syntactic similarity matrix. With N documents in our corpus, the syntax similarity matrix is $A_{N\times N}$; where $A_{i,j}$ is the syntactic similarity of the two documents i and j.

Next, spectral clustering (Shi & Malik, 2000) is used to cluster documents into a pre-defined number of groups. It has been shown that spectral clustering often outperforms traditional clustering algorithms (Von Luxburg, 2007). The general idea behind spectral clustering is to apply k-means clustering on eigenvectors of Laplacian matrix of A. The syntactic similarity matrix A, which is constructed in the previous step, and the number of clusters are provided as inputs to the spectral clustering method.

Clustering documents leads us to an essential next step which is extracting general attributes or representation of clusters. One way to address this concern is to calculate a centroid for each cluster. Clusters' centers facilitate researchers to better understand and analyze the syntactic structures used by a group of people or under certain situations by only analyzing center documents and without going through hundreds of documents. Hence, the third step in CASSIM-GR is calculating a centroid for each cluster. We define a cluster's center as the document which has the highest syntactic similarity to other documents in its cluster. To identify a cluster's center, we calculate average syntactic similarity of each document to other documents in its cluster and return the document with the highest average similarity. Additionally, we may return the top n documents with the highest average syntactic similarity to other documents in a cluster as representative samples of that cluster.

Finally, we use cross-validation to test the accuracy and representativeness of the clusters' centers. To cross-validate, our approach uses CASSIM to calculate the syntactic similarity of the left-out document to each centroid and assigns the document to a cluster with the highest similarity. This process is repeated N times and an accuracy of classification is reported by the method. In the following sections, we evaluate CASSIM-GR by performing classification experiments on three different corpora.

Experiments

We conducted three experiments to validate CASSIM-GR and to examine the representativeness of the cluster centroids. Additionally, we examined how well documents with similar syntactic structures cluster together and demonstrate the importance of syntactic similarity in classification. Further, we compare the accuracy of syntactic clustering to bag-of-words clustering. For this purpose, we use the tf-idf similarity matrix as input to spectral clustering. Lastly, we combined tf-idf and CASSIM-GR to see how including both sets of information affect the classification accuracy. Below, we discuss the three experiments in detail.

Experiment One

Experiment one was conducted on a corpus of syntactically similar documents. The corpus was generated by Amazon Mechanical Turk participants and consists of four groups of documents; each has high within-group syntactic similarity and low between-group syntactic similarity.

We used CASSIM-GR along with tf-idf, to group documents into clusters. Further, we combined these two approaches and calculated the overall accuracy. We first introduce the dataset and then report the results.

Data 118 MTurk participants answered a set of four questions. In each question they were asked to generate sentences with similar grammar rules to the sentence prompts in the question. Each of the four prompts had a different syntactic structure. Later, two independent coders, coded whether a sentence generated by a participant was grammatically similar to its prompt. Sentences which were identified as dissimilar by both coders were excluded from the dataset. Finally, a total of 272 documents, 68 documents in each group, were collected. See Boghrati et al. (2017) for more details.

Since participants were asked to write sentences similar to four different sets of prompts, the corpus is therefore divided to four separate groups, each associated to a question and its responses. Documents which are in the same group are considered to have similar syntactic structures.

Analysis We performed leave-one-out cross-validation for both of the clustering techniques. Namely, we ran the analysis on all the documents except for document i. Next, we labeled the clusters with the name of the group to which most of the documents belong. Then, we calculated similarity of document i to each cluster's center. Finally, document i was assigned to the cluster with which it had the highest syntactic similarity. The classification was considered successful if the assigned cluster's label and the document's group were identical.

We used the following approach to combine tf-idf and CASSIM-GR: First, we used CASSIM-GR and tf-idf approach separately to cluster documents into k clusters. Cluster $j, j \in [1,k]$ in tf-idf approach and cluster $j\prime, j\prime \in [1,k]$ in CASSIM-GR were labeled with the same name, that is, the majority of documents in cluster j and the majority of documents in cluster $j\prime$ were from the same group (e.g. 'liberals'). We averaged the syntactic similarity of document i to center of cluster j and the syntactic similarity of document i to center of cluster $j\prime$. We repeated this procedure k times to measure the similarity of document i to all k clusters and assigned document i to the cluster with highest similarity score. If the cluster's label and document i's label were the same, we would conclude that prediction was successful.

Results Our results demonstrate that CASSIM-GR is able to accurately cluster the corpus. Following the instructions discussed in above, we performed leave-one-out cross-validation on 272 documents. In each step, 271 documents were clustered in four groups and later the left-out document was assigned to one of the four clusters based on its similarity to the center of clusters.

Following this mechanism, CASSIM-GR yielded 95% accuracy while tf-idf approach was only 84.5% accurate. Running a chi-squared test demonstrates that CASSIM-GR results in significantly higher accuracy than tf-idf, $X^2(1) = 17.01, p < .001$. Since the dataset consists of groups of syntactically similar documents, it is not surprising that clustering based on syntactic structures surpasses the word-based approach and achieves a higher accuracy.

Next, we combined the two approaches and obtained an accuracy of 97.8%. While this result is not significantly higher than CASSIM-GR accuracy, $X^2(1) = 2.67, p = .10$, we may conclude that incorporating syntactic and semantic information together could potentially improve clustering accuracy.

Experiment Two

In the second experiment, we used the Dogmatism Dataset collected by Fast and Horvitz (2016). This dataset includes comments from New York Times which are rated based on their level of dogmatism. As explained below, we first categorized the documents as dogmatic or non-dogmatic based on this ratings. Next, we followed the procedure which was explained in the first experiment and clustered the documents using CASSIM-GR and the tf-idf approach. In the following subsections, we first introduce the dataset and then report the results.

Data The Dogmatism Dataset includes comments from New York Times. Amazon Mechanical Turk participants were asked to rate the level of dogmatism of each of the collected comments on a 5-point Likert scale. More details on the dataset and the annotation process are available at Fast and Horvitz (2016).

Analysis Dogmatism is subjective, and consequently inter-annotator agreement is higher for comments in both extreme sides of the spectrum. In other words, human coders tend to agree more on posts rated as very high in dogmatism and posts rated as very low in dogmatism (Fast & Horvitz, 2016). Following the method used by Fast and Horvitz (2016), to have a representative and balanced dataset, we selected the top 250 and the bottom 250 documents based on the dogma-

Table 1: Corpora Overview.

	Experiment One	Experiment Two	Experiment Three
Corpus	Syntactically Similar Sentences	Dogmatism in New York Times	Political Weblog Posts
Number of Groups	4	2	2
Number of Documents	272	500	452

Table 2: Accuracy of approaches in three experiments.

	Experiment One	Experiment Two	Experiment Three
CASSIM-GR	95%	54.8%	69.9%
TF-IDF Approach	84.5%	61%	64.4%
Combined Approach	97.8%	66.6%	71.9%

Table 3: Comparison of approaches in three experiments.

	Experiment One	Experiment Two	Experiment Three
CASSIM-GR vs. TF-IDF Approach	$X^2(1) = 17.01, p < .001$	$X^2(1) = 3.94, p < .05$	$X^2(1) = 3.13, p = .07$
TF-IDF Approach vs. Combined Approach	$X^2(1) = 29.61, p < .001$	$X^2(1) = 3.39, p = .06$	$X^2(1) = 5.89, p < .05$
CASSIM-GR vs. Combined Approach	$X^2(1) = 2.67, p = .10$	$X^2(1) = 14.59, p < .001$	$X^2(1) = .43, p = .51$

tism rating. We labeled the top 250 posts as dogmatic and the bottom 250 as non-dogmatic, hence the final dataset contained 500 posts with 250 in each group.

Results Following the instruction in Experiment 1, we performed leave-one-out cross-validation; we ran the clustering algorithm with 499 documents and left document $i, i \in [1,500]$, out. Then, we predicted to which cluster document i belonged. CASSIM-GR and tf-idf approach resulted in 55% and 61% accuracy respectively. Even though, the tf-idf approach outperformed our approach significantly, $X^2(1) = 3.94, p < .05$, combining these two approaches resulted in a higher accuracy of 66.6%, which is a marginally significant improvement over the tf-idf accuracy, $X^2(1) = 3.39, p = .06$.

This result provides evidence for the importance of syntactic structure similarity in clustering documents. It demonstrates that not only what different groups of people say, but also how they say what they say provide important information about the characteristics of the group. This is evident by the fact adding syntactic similarity to word-level similarity can improve the clustering accuracy.

Experiment Three

In this experiment, we applied CASSIM-GR on a corpus of political discussions taken from a set of conservative and liberal weblogs, and focus on the discussion about the Ground Zero Mosque (Dehghani, Sagae, et al., 2013).

Data The top five popular conservative and liberal news blogs were selected according to www.blogs.com. Next, a dataset of these weblogs posts which contained word *mosque* and were written in the time frame of the debate, were complied. For more details about the dataset and the data collection process please refer to Dehghani, Sagae, et al. (2013).

Analysis In this experiment, we randomly selected 250 posts from conservative weblogs posts and 250 posts from liberal weblogs posts, but due to encoding issues the final dataset included 226 posts from each group (total of 452 posts).

Results Similar to the previous experiments, we used the leave-one-out cross-validation procedure described above. Specifically, we trained the clustering algorithm on 451 documents and predicted to which cluster the left-out document belonged. This process was repeated 452 so that each document was tested once.

CASSIM-GR was able to successfully predict the correct cluster for a document with 70% accuracy, while tf-idf was 64.4% accurate. This difference is only marginally significant, $X^2(1) = 3.134, p = .0767$. Next, we combined these two approaches as described in the Experiment section. The total accuracy was 72% which is significantly more accurate than tf-idf approach alone, $X^2(1) = 5.8905, p = .0152$.

These results demonstrated that, in some cases, syntactic structures similarity may capture more crucial features needed for clustering compared to tf-idf approach. However, there are some features that only tf-idf approach can pick up. Thus, the combination of these two sets of features is needed for more accurate clustering.

Discussion and Future Work

Across three studies, we presented and validated a new approach called CASSIM-GR. CASSIM-GR clusters documents into separate groups based on their syntactic similarity and calculates a generalized representation of group-level syntax usage by performing four general steps: First, it creates a syntactic structure similarity matrix of documents using CASSIM. Second, it uses spectral clustering to group the documents into a pre-defined number of clusters using the syntactic similarity matrix generated in the previous step. Next, the algorithm selects the document which has the highest syntactic similarity to the other documents within each cluster and identifies it as the centroid of that cluster. Finally, it can be used to classify unknown documents based on the document's syntactic similarity to the clusters' centers.

We applied CASSIM-GR to three unique corpora (Table 1) across three experiments to compare its accuracy to both a bag-of-words approach and a combined approach incorporating tf-idf semantic information and CASSIM-GR. As Table 2 demonstrates, tf-idf and CASSIM-GR varied in their relative strength for clustering accuracy across studies. The combined approach incorporating both syntactic (CASSIM-GR) and semantic (tf-idf) information resulted in the highest clustering accuracy across all three experiments. While not a significant improvement beyond both single approaches, the combination approach significantly outperformed tf-idf in two of the three experiments and CASSIM-GR in the second experiment. Therefore, we may conclude that word-level similarity and syntactic similarity capture different aspects of language, and consequently, combining the two features' similarities results in more accurate clusters.

Our results indicate that methods assessing syntactic similarity may more accurately cluster documents than methods which rely on semantics alone. While there may be situations in which groups use the same general words to discuss a topic, syntactic similarity differences could still allow researchers to distinguish between different subsets of individuals.

More importantly, CASSIM-GR gives researchers an opportunity to study syntactic differences between groups by analyzing the prototypical syntactic structures at the clusters' centers. The syntactic structures used by a cluster's center document is defined as a generalized representation of syntactic structures of the documents in that cluster. Assessing differences in these structures may help to capture underlying psychological differences between groups in the ways that they conceptualize a topic or how they communicate with each other.

A vital component of CASSIM-GR is measuring syntactic similarity among documents using CASSIM. As mentioned previously, CASSIM's general focus is on comparing constituency parse trees. Building on CASSIM, we intend to compare dependency parse trees among sentences and documents to add another syntactic similarity measurement to CASSIM. Unlike constituency parse trees which posit the connection between part of speech tags, dependency parse trees reveals the relationship between the words in a sentence. By incorporating this feature into CASSIM, researchers may further use CASSIM-GR not only to generalize syntactic structure of a group of documents, but also their dependency structures. This extension will help researchers study human language in finer grained detail by looking at the relationship between words.

In summary, we introduced a new method for computing generalized representations of syntactic structures of documents, allowing researchers to distinguish between groups of documents based on syntactic differences. Further, In the three experiments, we demonstrated the benefits of including syntactic structure similarity scores in clustering documents. In each experiment, we repeated a clustering procedure, once using CASSIM-GR and once using tf-idf similarity matrix. Then, we calculated clustering accuracy of each approach using leave-one-out cross-validation mechanism. Finally, we combined the results of these two approaches and calculated the accuracy when both sets of features were present. Our results support our assumption and demonstrated that syntactic similarity scores capture different aspects of language compared to bag-of-words, and therefore help improve clustering accuracy.

Acknowledgments

This research was supported in part by NSF IBSS Grant #1520031.

References

Boghrati, R., Hoover, J., Johnson, K. M., Garten, J., & Dehghani, M. (2017). Conversation level syntax similarity metric. *Journal of Behavior Research Methods*.

Bresnan, J., & Hay, J. (2008). Gradient grammar: An effect of animacy on the syntax of give in new zealand and american english. *Lingua*, *118*(2), 245–259.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. In *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510–516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dehghani, M., Bang, M., Medin, D., Marin, A., Leddon, E., & Waxman, S. (2013). Epistemologies in the text of children's books: Native-and non-native-authored books. *International Journal of Science Education*, *35*(13), 2133–2151.

Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., ... Graham, J. (2016). Purity homophily in social networks. *Journal of Experimental Psychology: General*.

Dehghani, M., Sagae, K., Sachdeva, S., & Gratch, J. (2013). Linguistic analysis of the debate over the construction of the ground zero mosque. *Journal of Information Technology & Politics. Advance online publication. doi*, *10*(19331681.2013), 826613.

Dehghani, M., Sagae, K., Sachdeva, S., & Gratch, J. (2014). Analyzing political rhetoric in conservative and liberal

weblogs related to the construction of the ground zero mosque. *Journal of Information Technology & Politics*, *11*(1), 1–14.

Fast, E., & Horvitz, E. (2016). Identifying dogmatism in social media: Signals and models. *arXiv preprint arXiv:1609.00425*.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Gawda, B. (2010). Syntax of emotional narratives of persons diagnosed with antisocial personality. *Journal of psycholinguistic research*, *39*(4), 273–283.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193–202.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029.

Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, *6*, 287–317.

Jahr, E. H. (1992). Middle-aged male syntax. *International Journal of the Sociology of Language*, *94*(1), 123–134.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786.

Lewis, D. D. (1992). Feature selection and feature extraction for text categorization. In *Proceedings of the workshop on speech and natural language* (pp. 212–217).

Lewis, D. D., & Croft, W. B. (1989). Term clustering of syntactic phrases. In *Proceedings of the 13th annual international acm sigir conference on research and development in information retrieval* (pp. 385–404).

Lin, Y.-S., Jiang, J.-Y., & Lee, S.-J. (2014). A similarity measure for text classification and clustering. *IEEE transactions on knowledge and data engineering*, *26*(7), 1575–1590.

Liu, L., Kang, J., Yu, J., & Wang, Z. (2005). A comparative study on unsupervised feature selection methods for text clustering. In *Natural language processing and knowledge engineering, 2005. ieee nlp-ke'05. proceedings of 2005 ieee international conference on* (pp. 597–601).

Liu, T., Liu, S., Chen, Z., & Ma, W.-Y. (2003). An evaluation on feature selection for text clustering. In *Icml* (Vol. 3, pp. 488–495).

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496.

Maass, A., Karasawa, M., Politi, F., & Suga, S. (2006). Do verbs and adjectives play different roles in different cultures? a cross-linguistic analysis of person representation. *Journal of personality and social psychology*, *90*(5), 734.

Matlock, T. (2001). *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.

Mehl, M. R., Robbins, M. L., & Holleran, S. E. (2012). How taking a word for a word can be problematic: Context-dependent linguistic markers of extraversion and neuroticism. *Journal of Methods and Measurement in the Social Sciences*, *3*(2), 30–50.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. No. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.

Ramirez-Esparza, N., Chung, C. K., Kacewicz, E., & Pennebaker, J. W. (2008). The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In *Icwsm.*

Sasaki, M., & Shinnou, H. (2005). Spam detection using text clustering. In *Cyberworlds, 2005. international conference on* (pp. 4–pp).

Sedding, J., & Kazakov, D. (2004). Wordnet-based text document clustering. In *proceedings of the 3rd workshop on robust methods in analysis of natural language data* (pp. 104–113).

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(8), 888–905.

Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Song, W., Li, C. H., & Park, S. C. (2009). Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures. *Expert Systems with Applications*, *36*(5), 9095–9104.

Vigliocco, G., & Franck, J. (1999). When sex and syntax go hand in hand: Gender agreement in language production. *Journal of Memory and Language*, *40*(4), 455–478.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*(4), 395–416.

Zheng, H.-T., Kang, B.-Y., & Kim, H.-G. (2009). Exploiting noun phrases and semantic relationships for text document clustering. *Information Sciences*, *179*(13), 2249–2262.