PSYCHONET
A Psycholinguistc Commonsense Ontology

Haytham Mohtasseb and Amr Ahmed
School of Computer Science, University of Lincoln, Brayford Pool, Lincoln, U.K.
{*hmohtasseb, aahmed*}*@lincoln.ac.uk*

Keywords: Commonsense knowledgebase, Semantic network, Ontology development, Psycholinguistic, Text classification.

Abstract: Ontologies have been widely accepted as the most advanced knowledge representation model. This paper introduces PsychoNet, a new knowledgebase that forms the link between psycholinguistic taxonomy, existing in LIWC, and its semantic textual representation in the form of commonsense semantic ontology, represented by ConceptNet. The integration of LIWC and ConceptNet and the added functionalities facilitate employing ConceptNet in psycholinguistic studies. Furthermore, it simplifies utilization of the huge network of ConceptNet for a specific multimedia application based on key category(ies) from LIWC, such as visual or biological applications. PsychoNet adds a new layer of complementary psycholinguistic functions to the original semantic network. Moreover, learning, either clustering or classification, is more applicable in the developed ontology. The paper shows a sample application of text classification for mood prediction task. The result confirms the validity of the proposed network as PsychoNet outperforms LIWC in mood prediction.

1 INTRODUCTION

The considerable development of multimedia communication goes along with an exponentially increasing volume of textual information. Ontologies have been widely accepted as the most advanced knowledge representation model. They are a very crucial part of information extraction, semantic web, knowledge discovery, and computational linguistic. Huge effort is needed from the domain expert in order to construct ontologies manually. There is a need for automatic approaches in ontology building which will help the domain experts in constructing extensive domain ontologies efficiently.

The ontology engineering community convene to develop more works toward integrating ontologies so that they can share and reuse each others knowledge (Noy and Hafner, 1997). If one ontology, for example, has a well-developed theory of psychology, another ontology (say, the one representing commonsense experiments) could then use this theory without having to reinvent it. We propose the use of psycholinguistic lexicon in order to find groups of concepts which are related to each other. Such groups of related concepts will enable the domain expert to either, evaluate and update the existing ontology in case those concepts are already defined in the ontology, or to enrich the existing ontology in case those concepts are not defined.

This paper introduces a novel commonsense knowledgebase that forms the link between the psycholinguistic and its semantic textual representation. We refer to it as "PsychoNet". This knowledgebase is built by a fully automated engine that performs lexical analysis on concepts and extracts the corresponding psycholinguistic categories. It allows the researcher to use one coherent knowledgebase that has the power of semantic commonsense and psycholinguistic taxonomy.

There are many types of tagging/integration, but this study presents the benefits of integrating LIWC and ConceptNet for many applications. This paper develops ConceptNet, a commonsense ontology (Liu and Singh, 2004), by adding a psycholinguistic layer, utilizing LIWC (Pennebaker et al., 2001), on the top of ontology.

The rest of the paper is organized as following. In section 2, we review the recent work related to our domain. Section 3 shows our work that starts by presenting the structure of PsychoNet, introducing the new functions, and finally illustrating an application of text classification using PsychoNet. Finally, we show the conclusions and our future work.

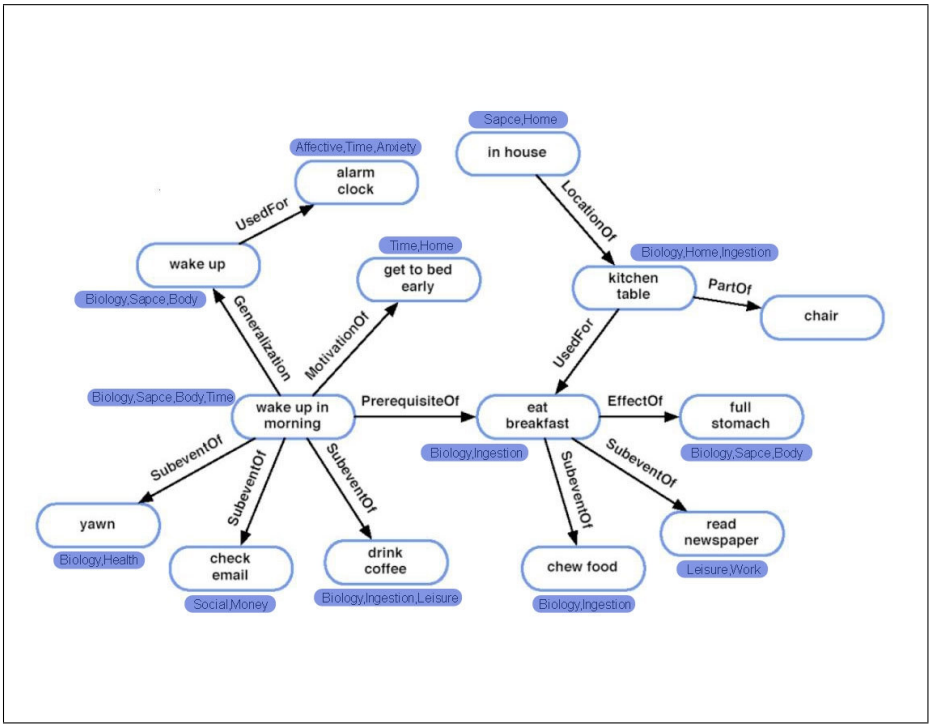

Figure 1: PsychoNet.

2 BACKGROUND

2.1 LIWC

Linguistic Inquiry Words Count (LIWC) (Pennebaker et al., 2001) was constructed by having groups of judges evaluate the degree to which about 2000 words or word stems were related to each of several dozen categories. The categories include negative emotion words (sad, angry), positive emotion words (happy, laugh), standard function word categories (first, second, and third person pronouns, articles, prepositions), and various content categories (e.g., religion, death, occupation). LIWC computes the percentage of total words that these and other linguistic categories represent (Chung and Pennebaker, 2007).

LIWC has been extensively validated and has provided substantial evidence about the social and psychological implications of word use (Pennebaker et al., 2003). The selected 63 LIWC features are grouped into four types:

1. Standard linguistic features (e.g., total word count, word per sentence, pronouns, punctuations, articles, time);
2. Psychological features (e.g., affect, cognition, biological processes);
3. Personal concerns features (e.g., work, sports, religion, sexuality);
4. Paralinguistic features *assents* (e.g., agrees, ok), *fillers* (e.g., err, umm), *non fluencies* (e.g., I mean, you know).

LIWC can handle the different stems of the word, which is one of the common issues in natural language processing NLP. So the stem *hungr* captures the words *hungry, hungrier, hungriest* and so on dictionary.

2.2 ConceptNet

ConceptNet is currently considered to be the largest commonsense knowledgebase (Liu and Singh, 2004). The Open Mind commonsense knowledgebase has been analyzed to create ConceptNet, a large semantic network currently containing over 250,000 nodes. Nodes are semi-structured English fragments, interrelated by an ontology of twenty semantic relations (predicates). The predicates are machine-readable of the form: (IsA ”tennis” ”sport”) and (EventForGoalEvent ”play tennis” ”have racket”).

Each node is a concept, which is a part of a sentence that expresses a meaning. ConceptNet is a very rich knowledgebase for several aspects: First, the

huge number of assertions and nodes contained. Second, the wide range of information included. Finally, the various types of relationships that hold description parameters existed. ConceptNet is very useful in describing real life scenes that make it a good candidate to be integrated with LIWC.

2.3 More Nets

Many developments over ConceptNet had been implemented to create adapted semantic networks. LifeNet is created utilizing the temporal relations from ConceptNet citation (Singh et al., 2004). This network adds a variety of temporal operations like predicting what else might be true now, in the near future or in the near past, explaining why some events have happened, or filtering nodes that are not likely to be true. Moreover, EventNet used the temporal nodes in LifeNet to create an association network (Espinosa and Lieberman, 2005). It can make predictions of the more likely previous or following events associated with a certain set of events. In additions, (Altadmri and Ahmed, 2009) proposed VisualNet as a novel commonsense knowledgebase that forms the link between the visual world and its semantic textual representation. VisualNet is obtained by constructing a unified structure concluding the knowledge from WordNet and ConceptNet. To the best of our knowledge, this paper introduces the first development of ConceptNet towards psycholinguistic direction. PsychoNet develops both ConceptNet and LIWC. It enriches LIWC by adding the semantic dimension to its content and representing the psycholinguistic categories using commonsense concepts rather than words. In other words, LIWC users can query the taxonomy using contextual concepts instead of terms. On the other hand, PsychoNet simplifies text classification in ConceptNet and allows filtering the huge concept graphs based on a key category for a specific application. The next section will explain in details the characteristics of PsychoNet.

3 PSYCHO NET

The node in PsychoNet is a concept associated with a psychometric field that contains the psycholinguistic categories and their relevance degree. Figure 1 shows a snapshot of PsychoNet describing various activities of everyday morning. We can see that "Biology" is the main theme of the graph as the majority of nodes outline eating, drinking, and ingesting activities. The graph also highlights other indications about the place which is at "Home". PsychoNet makes the graph easily understood by human (very fast to read what the main theme is). From now, we would refer PNet to PsychoNet and CNet to ConceptNet. PNet can be built through the following 3 stages:

- Concept Psycho-annotation: Add matching LIWC categories and frequencies to each node in CNet.
- Predicate Psycho-annotation: Use the dominant psycho-category within PNet nodes.
- Cleaning: Deprecate the concepts and predicates that do not have matching psycho-category.

Both LIWC and CNet have been improved in PNet representation. The content of LIWC dictionary is fixed as there are specific words for each category. PNet creates a new representation of LIWC based on concepts rather than words. Although the two representations seem similar, as concepts compound from words, but in fact they are different. The concept is consisted from words mentioned in context to form a meaningful thing. However, individual terms not always give full meaning and have some ambiguity. Furthermore, the semantic network allows expanding the categories by including new words using the relational predicates resulting in a new semantic level of knowledge added over LIWC. The functions of the ontology like Get-Topic and Get-Context summarized by LIWC categories would help the researchers in psycholinguistic field.

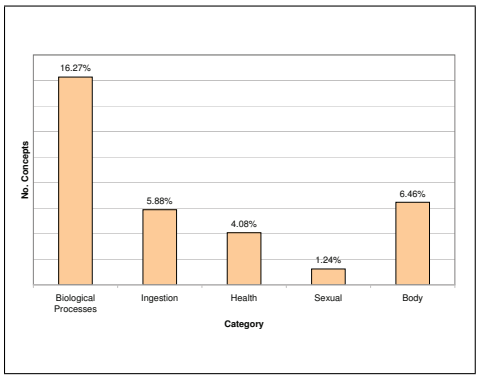

Figure 2: Concepts spread for Biocomputing applications.

On the other hand, the main benefit brought to CNet is to deal smoothly with more than 250,000 concepts. Although there are many functions like Get-topic, Get-Context, and Get-analogy; having the high-level LIWC categories provide a new mechanism to navigate through the network. About 70% of the concepts have their corresponding LIWC categories. The psycholinguistic categories provide the ability to subscan the network using a key category focusing on a specific task. For example, Figure 2 shows the spread

of biology-related concepts that could be utilized in biocomputing applications. Moreover, Video retrieval applications (Altadmri and Ahmed, 2009) would improve their performance by reducing the size of network and concentrating on video concepts as depicted in Figure 3.

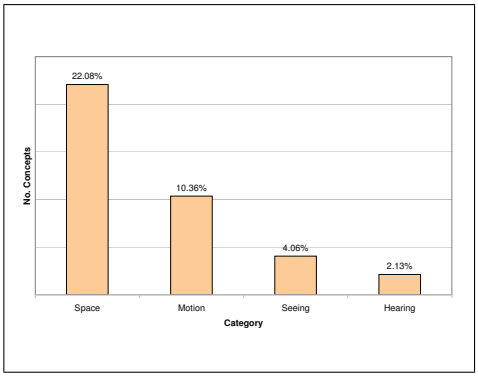

Figure 3: Concepts spread for video mining applications.

3.1 New Functions

CNet provides several functions over its semantic network such as GuessTopic, TopicGisting, and GuessMood. But, PNet adds novel functions that improve the usability of CNet in many applications. The rest of this section presents the description of each function.

Emotional Degree. The emotional-degree function is calculated as the difference between the LIWC scores for the concepts belonging to positive emotion concepts (e.g., happy, good, nice) and negative emotion category (e.g., kill, ugly, guilty). Higher scores indicate greater overall positive emotion. Emotional-degree function gives the overall emotional sense, while GuessMood return the emotional sense based on different six moods. The new function is useful to get an overall single value or binary emotion. However, for more detailed emotion, it can be accompanied with GuessMood result.

Social Orientation. The social-orientation function indicates how often users used words such as talk, share, or friends and personal pronouns other than first-person singular (Cohn et al., 2004). Psychologically, it reflects the personality of users as being extroverts or introverts.

Psycholinguistic Index. The psycholinguistic-index function gives the overall psycholinguistic summary of the intended semantic graph. It converts the graphs to a numerical vector in which the cells represent the weighting balance of each LIWC category like social, biology, or cognition.

Psychometric Similarity. The psychometric-similarity function measures the similarity degree between semantic graphs based on the psycholinguistic distance between concepts. Cosine distance is utilized across the psycholinguistic vectorized representation of the two graphs. Mainly, this function is useful for clustering applications.

3.2 Mood Classification

In this section, we present a sample application of using PNet in text classification. The main contribution is in the improvement in accuracy achieved using PNet compared to LIWC. Figure 4 and 5 show the required stages for building a classification model distinguishing between moods using LIWC and PNet respectively. The difference between the two experiments is how the learning vectors have been created either from words (figure 4) or by applying psycholinguistic-index function over concepts (figure 5).

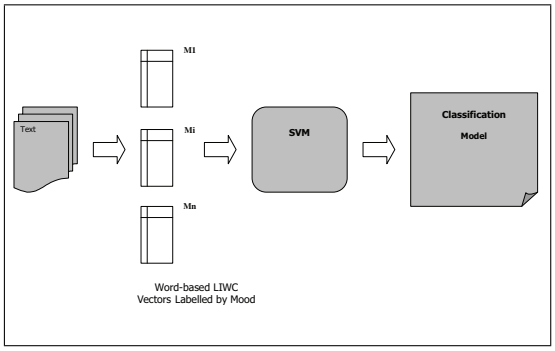

Figure 4: Mood classification using LIWC.

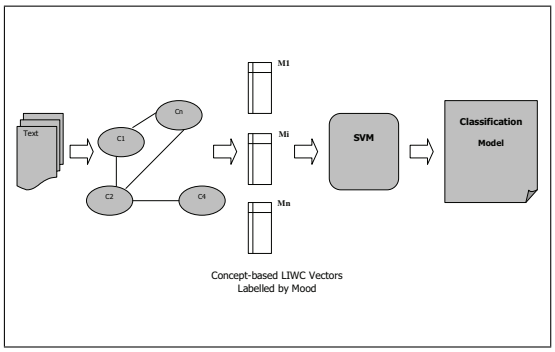

Figure 5: Mood classification using PNet.

3.2.1 Corpus

We selected as a corpus one of the famous personal blog sites, namely "LiveJournal"[1]. LiveJournal is a free personal blog website forming a community on the Internet that contains millions of users publishing their own ongoing personal diaries. We downloaded from LiveJournal 21,000 blog posts for various moods. Bloggers in Livejournal are given the facility to tag their blog post with an optional field indicating the "current mood" which we use it as the ground-truth.

3.2.2 Experiment

The blog posts are converted to numerical vectors in which the entries contain the corresponding features values. The next step after moving to feature space is using machine learning. We choose Support Vector Machines (SVM) as the classification algorithm which is one of the best algorithms in this domain. For each mood, random training and testing sets have been constructed from the set of posts labeled with that mood as positive examples, and an equal amount of negative examples, form all other moods. Since many moods did not have large amounts of associated blog posts, the experiment is limited to report the results for most frequent ten moods. For each mood, we have the following classification contingency table:

Table 1: Classification contingency table.

		Real Value	
		Yes	No
Classifier Value	Yes	TP	FP
	No	FN	TN

TP (True Positive) is the correctly classified instances, TN (True Negative) is the correctly rejected instances, FP (False Positive) is the incorrectly classified instances, FN (False Negative) is the incorrectly rejected instances. Based on the contingency table, the following standard classification measures are defined:

1. Precision
$$= \frac{TP}{TP+FP}$$
2. Recall
$$= \frac{TP}{TP+FN}$$
3. F-Measure
$$= \frac{2.Precision.Recall}{Precision+Recall}$$

[1] http://www.livejournal.com

Table 2 shows the results using the three above defined measures: Precision, Recall, and F-Measure. Generally, PNet outperforms LIWC for most of the moods (significant results are in bold). The next section shows a more detailed discussion of the results.

3.2.3 Discussion

When LIWC alone has been tried in mood classification task, the results were poor and not promising. LIWC had been used successfully in numerous text analyses tasks for analyzing the emotions of users in blog text (Gill et al., 2008; Hancock et al., 2008; Hancock et al., 2007), identifying the gender of bloggers (Nowson and Oberlander, 2006), recognizing the personality (Gill, 2003; Mairesse et al., 2007), and for author identification (Mohtasseb and Ahmed, 2009a; Mohtasseb and Ahmed, 2009b).

The target classes (Gender, Age, User ID) in the previous mentioned text classification tasks, where LIWC produced good results, are actually facts. However, in mood classification, the target class (mood) is provided by user. So it is subjective rather than objective data. It is usual that a user tag some posts with different moods even where the contents are, to some extent, similar.

Hence, this task is challenging and LIWC features alone can not fulfill the task. Previous studies in mood prediction confirm this difficulty as they utilized various types of features in order to achieve reasonable results (Mishne, 2005; Leshed, 2006). Using PNet improves the result of mood classification over LIWC. This is performed by only picking the concepts and producing the summarized LIWC vector of the extracted concepts. PNet enhanced the result for some moods and improved accuracy to above 50% for others. Althought the resulting accuracy may not be higher than what is reported in literature (60%) we should emphasize that this results is based on PNet only, in comparison with LIWC only. Hence, it is highly expected that adding all other features (as in literature) will result in better overall accuracy. This puts PNet up as a candidate features set to be included with other well proved features to contribute in mood attribution task.

4 CONCLUSIONS

In this paper, we introduced a novel commonsense knowledgebase, PsychoNet, for high-level psycholinguistic semantic domain applications. The proposed knowledgebase manages to merge advantages and functionalities of both LIWC and ConceptNet. The

Table 2: Classification result.

Mood	Recall		Precision		F-Measure	
	PNet	LIWC	PNet	LIWC	PNet	LIWC
amused	**0.58**	0.46	**0.54**	0.35	**0.56**	0.40
cheerful	**0.48**	0.37	**0.48**	0.40	**0.48**	0.39
busy	**0.50**	0.34	**0.64**	0.49	**0.56**	0.40
happy	**0.52**	0.42	**0.59**	0.41	**0.56**	0.42
calm	**0.50**	0.34	**0.39**	0.32	**0.44**	0.33
content	**0.41**	0.29	**0.42**	0.27	**0.42**	0.28
creative	0.30	0.43	0.20	0.31	0.24	0.36
bored	**0.53**	0.41	**0.47**	0.38	**0.50**	0.39
contemplative	**0.46**	0.42	**0.44**	0.24	**0.45**	0.30
exhausted	0.31	0.43	0.28	0.45	0.30	0.44

new annotation of nodes in PsychoNet makes its usage easier in many text analysis areas such as information extraction, semantic web, and text mining. An experiment on a sample application, which is mood classification based on the proposed knowledgebase has been demonstrated showing the improvement of PsychoNet over LIWC for several moods.

Traditional text mining techniques tend to summarize too much irrelevant information as a term can have different meanings in distinct contexts. However, the proposed method that is based on ontological concepts is more effective as they avoid such ambiguity. PsychoNet adds novel functions that improve the usability of ConceptNet in many applications such as biocomputing and video mining. This paper opens new research directions by introducing a psycho-ontology to psycholinguistic studies.

REFERENCES

Altadmri, A. and Ahmed, A. (2009). Visualnet: commonsense knowledgebase for video and image indexing and retrieval application. In *IEEE International Conference on Intelligent Computing and Intelligent Systems, 2009. ICIS 2009*, volume 3.

Chung, C. K. and Pennebaker, J. W. (2007). The psychological function of function words. *Social communication: Frontiers of social psychology*, pages 343–359.

Cohn, M. A., Mehl, M. R., and Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding september 11, 2001. *Psychological Science*, 15(10):687–693.

Espinosa, J. and Lieberman, H. (2005). Eventnet: Inferring temporal relations between commonsense events. *MICAI : Advances in Artificial Intelligence*, pages 61–69.

Gill, A. (2003). Personality and language: The projection and perception of personality in computer-mediated communication.

Gill, A. J., French, R. M., Gergle, D., and Oberlander, J. (2008). The language of emotion in short blog texts. In *Proceedings of the ACM 2008 conference on Computer supported cooperative work*, pages 299–302. ACM New York, NY, USA.

Hancock, J. T., Gee, K., Ciaccio, K., and Lin, J. M. H. (2008). I'm sad you're sad: emotional contagion in cmc. In *Proceedings of the ACM 2008 conference on Computer supported cooperative work*, pages 295–298. ACM New York, NY, USA.

Hancock, J. T., Landrigan, C., and Silver, C. (2007). Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 929–932. ACM New York, NY, USA.

Leshed, G. (2006). Understanding how bloggers feel: recognizing affect in blog posts. In *Conference on Human Factors in Computing Systems*, pages 1019–1024. ACM New York, NY, USA.

Liu, H. and Singh, P. (2004). Conceptneta practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500.

Mishne, G. (2005). Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*.

Mohtasseb, H. and Ahmed, A. (2009a). Mining online diaries for blogger identification. In *The 2009 International Conference of Data Mining and Knowledge Engineering (ICDMKE'09)*.

Mohtasseb, H. and Ahmed, A. (2009b). More blogging features for author identification. In *The 2009 International Conference on Knowledge Discovery (ICKD'09)*.

Nowson, S. and Oberlander, J. (2006). The identity of bloggers: Openness and gender in personal weblogs. In *Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs*.

Noy, N. F. and Hafner, C. D. (1997). The state of the art in ontology design. *AI magazine*, 18(3):53–74.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway : Lawrence Erlbaum Associates*.

Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577.

Singh, P., Barry, B., and Liu, H. (2004). Teaching machines about everyday life. *BT Technology Journal*, 22(4):227–240.