PSYCHONET 2
Contextualized and Enriched Psycholinguistic Commonsense Ontology

Haytham Mohtasseb, Amr Ahmed, Amjad AlTadmri and David Cobham
School of Computer Science, University of Lincoln, Brayford Pool, Lincoln, U.K.

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

Abstract: PsychoNet 1 has demonstrated the feasibility of integrating psycholinguistic taxonomy, represented in LIWC, and its semantic textual representation in the form of commonsense ontology, represented in ConceptNet. However, various limitations exist in PsychoNet 1, including the lack of concluding context of the concept annotation. In this paper, we address most of those limitations and introduce a new enhanced and enriched version, PsychoNet 2. PsychoNet 2 utilizes WordNet, in addition to LIWC and ConceptNet, to produce an integrated contextualized psycholinguistic ontology. The first and the main contribution is that, in PsychoNet 2, each concept is annotated by the potential (most representative) contextual psycholinguistic categories, rather than all applicable categories. The second contribution is the enrichment of LIWC through utilizing WordNet. This in fact produced an enriched version of LIWC that may also be used independently in other applications. This has contributed to substantial enrichment of PsychoNet 2 as it facilitated including additional number of concepts that were not included in PsychoNet 1 due to lack of corresponding words in the original LIWC. A sample application of text classification, for a mood prediction task, is presented to demonstrate the introduced enhancements. The results confirm the improved performance of the new PsychoNet 2 against PsychoNet 1.

1 INTRODUCTION

The ontology engineering community is increasingly convening to develop more work towards integrating ontologies so that they can share and reuse each other's knowledge (Noy and Hafner, 1997). PsychoNet 1 (Mohtasseb and Ahmed, 2010b) introduced a novel commonsense knowledgebase that forms the link between the psycholinguistic and its semantic textual representation. 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 and integration (more details in Section 2), but this study presents the benefits of integrating LIWC, ConceptNet, and WordNet for a wide range of applications. This paper develops ConceptNet, a commonsense ontology (Liu and Singh, 2004), by adding a psycholinguistic layer, utilizing LIWC (Pennebaker et al., 2001), enriched by the lexical semantic network namely WordNet (Miller, 1995). Furthermore, in PsychoNet 2, only the common highly rated annotations are kept as they represent the context of the concept.

The rest of the paper is organized as follows. Section 2 reviews the recent work related to our domain. Section 3 presents PsychoNet 2 including the enrichment and the contextualization processes. Section 4 shows the application of PsychoNet 2 in mood classification and its results. Finally, the paper is concluded in Section 5.

2 BACKGROUND

This section presents an overview of the related work and the existing development in the same area including LIWC, ConceptNet, WordNet, and PsychoNet 1.

Linguistic Inquiry Words Count (LIWC) (Pennebaker et al., 2001) has been built by classifying a nominated set of 2000 words (and word stems) into several dozens of psycho categories, based on the judgment of a group of linguistic experts. The categories include positive and negative emotional words , functional words (pronouns, articles, prepositions), health and biology categories, and other contextual categories (e.g., sport, family, religion, death). LIWC had been used successfully in numerous text analyses tasks for analyzing the emotions of users in blog

Mohtasseb H., Ahmed A., AlTadmri A. and Cobham D..
PSYCHONET 2 - Contextualized and Enriched Psycholinguistic Commonsense Ontology.
DOI: 10.5220/0003621403390343
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2011), pages 339-343
ISBN: 978-989-8425-80-5

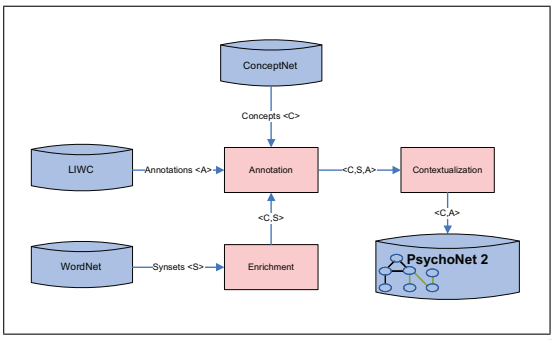

Figure 1: PsychoNet 2 Building Framework.

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), studying the demographic differentiations across the styles of bloggers (Mohtasseb and Ahmed, 2010a), and in authorship identification (Mohtasseb and Ahmed, 2009a; Mohtasseb and Ahmed, 2009b). However, all of these tasks have been applied on the word level rather than the concept level, which is available in PsychoNet 1.

The ConceptNet knowledgebase is a semantic network encompasses the spatial, physical, social, temporal aspects of everyday life (Liu and Singh, 2004). ConceptNet is generated automatically from the 700,000 sentences of the Open Mind Common Sense corpus[1]. ConceptNet is currently considered to be the largest commonsense semantic network containing over 250,000 nodes. Nodes are semi-structured English fragments, interrelated by an ontology of twenty semantic relations (predicates). ConceptNet is very useful in describing real life scenes which makes it a good candidate to be integrated with LIWC that will add the psycholinguistic dimension.

WordNet is a large lexical database of English (Miller, 1995). Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets). It is a very rich domain-independent knowledgebase of lexical units that consist of various forms of synonyms. WordNet is effective for studying the relationships within similar words in terms of meaning, generalization or specialization.

On the other hand, PsychoNet 1 introduced the first development of ConceptNet towards psycholinguistic direction, utilizing LIWC. It has been 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. Moreover, PsychoNet 1 simplified applying text classification tasks in ConceptNet and allows filtering the huge concept graphs based on a key category for a specific application. PsychoNet 2 introduces further improvement on PsychoNet 1 as being explained in the next section.

[1]http://web.media.mit.edu/ push/OMCS-Research.html

3 PsychoNet 2

In PsychoNet 1 (Mohtasseb and Ahmed, 2010b), each node is a concept associated with a psychometric field that contains the psycholinguistic categories (annotations) and their relevance degree. In PsychoNet 2, many limitations have been addressed including missing concepts and contextualization, and more substantial improvements are introduced through the addition of two new stages as depicted in Figure 1. The first stage, *Enrichment*, utilizes WordNet to deal with those concepts, existing in ConceptNet, that do not have matching LIWC annotations. The resulting synonym sets, for the original component words, are then annotated using LIWC. This is explained in detail in Section 3.1. Section 3.2 presents the second stage, *Contextualization*, that starts by selecting the synonym sets that share the same set of annotations. Then, it deduces the high ranked annotations that potentially represent the context of the concept. The following subsections explain the two new stages; Enrichment and Contextualization, respectively.

3.1 Enrichment

Through our analysis of PsychoNet 1, it has been found that there were 21498 concepts that have not been included. Moreover, the analysis showed that 31863 words, which belong to the commonsense concepts, do not have matching LIWC categories. To address this and try to annotate and include most concepts, we had to develop a way to enrich LIWC to include those missing words and their variations. Therefore, WordNet is utilized here to expand and enrich the contents of LIWC based on the commonsense words of ConceptNet, as explained below.

Assume that $W = \{w_1, w_2, \ldots, w_n\}$ is the set of commonsense words that do not have LIWC annotations. For each word $w_i \in W$, all synsets (synonym sets) $\{S_1, S_2, \ldots, S_m\}$, of this word, are extracted using WordNet. Hence, $S_j = \{s_1, s_2, \ldots, s_l\}$ represents one of the synsets where s_k is a synonym for w_i within the context of the synset S_j. $A_{S_j} = \{a_1, a_2, \ldots, a_z\}$ is the list of LIWC annotation of S_j if there were cross joint annotations across all s_k. Then, the set of final LIWC

Table 1: Snapshot of the result showing the added commonsense words to LIWC using WordNet.

Word	Synset	Annotation
earth	world,globe	Relativity,Space
earth	ground	Relativity,Space
absorbing	engross,engage, occupy	Affective, Pos.Emotion
live	alive	Biological, Health,Death
live	exist,survive, subsist	Achievement
awake	alert,alive	Biological, Health
cereal	food grain	Ingestion, Biological
newspaper	paper	Work
audience	hearing	Perceptual, Social,Hearing
calculate	account	Money
gift	endow,empower, invest,endue	Money
gift	endowment,talent	Affective, Pos.Emotion
crime	law-breaking, offense	Neg.Emotion, Affective,Anger

annotations A_i of w_i is produced by the union of the annotations of synsets $A_i = \{A_{S_1} \cup A_{S_2} \cup \cdots \cup A_{S_v}\}$.

According to the approach described above, if a word w_i has a non-empty annotation set A_i, then w_i is added to the corresponding list of words of its relevant psycholinguistic categories. In addition, the annotation list A_i will contribute to the concept annotation where w_i originated from (Section 3.2).

Table 1 shows a snapshot of the resulting new words along with the assigned annotations. As a result of the above enrichment stage, 7772 words have been added to LIWC, 8663 new concepts have been included in PsychoNet 2, and 56615 concepts have been enriched with extra annotations. This is a mutual benefit for those who want to use LIWC alone, with this enriched version, and for those who still need to use the full PsychoNet 2 knowledgebase.

It is worth mentioning that the number of annotation sets A_{S_j} might not be equal to the number of synsets. This is because in some cases there might be a synset S_j that has no representative psycholinguistic annotation (i.e. has an empty A_{S_j} annotation set). Therefore, we can see in Table 1 that the enrichment process provides two matching synsets with different sets of annotation for the word *live*, however, it only provides one annotation for the word *newspaper*.

3.2 Contextualization

PsychoNet 1 associates each concept with a list of psycholinguistic annotations and its corresponding frequencies. This is due to the existence of common annotations across the words of the concept. Although there could be multiple annotations for the same word, it should only select the annotations that are related to the context. In PsychoNet 2, it is intended to select the psycholinguistic annotations based on the context of the representing words. This will maintain only psycholinguistic annotations (categories) which suit the context of the concept. Table 2 shows an example of annotations results before and after contextualization.

We can see that the concept *"a scream of freedom"* has conflicting annotations; *Neg.Emotion* and *Pos.Emotion*, resulting from its component words. Moreover, it located *Hearing* which is related to one of the words, but it is outside of the context for this concept. The proposed algorithm ended up with *Affective* annotation which is more representative of the context of that concept. The same can be seen in the concept *"The best way to commit a crime"*. Similarly, the concept *"coffee shop"* has *Leisure* as the context annotation. Many other concepts have not been included before in PsychoNet 1, such as *"hit ball"* and *"zip code"*. But now, in PsychoNet 2, they are included and annotated, thanks to the enrichment of the LIWC by utilizing the WordNet (Section 3.1).

4 APPLICATION: MOOD CLASSIFICATION

This section presents a sample text classification application using PsychoNet 2. The contribution lies in accuracy improvement achieved using PsychoNet 2 compared to PsychoNet 1 and LIWC respectively. We utilized the same mood experiment framework and corpus presented in (Mohtasseb and Ahmed, 2010b) for building a classification model distinguishing between moods using LIWC and PsychoNet, for both versions, respectively. The difference between the two experiments derives from creating the learning vectors either by using LIWC to extract the features from words, or by applying psycholinguistic-index function (Mohtasseb and Ahmed, 2010b) over the extracted concepts. For each mood, the F-Measure value of the classification result is calculated. Results presented in table 3 shows that PsychoNet 2 outperforms both LIWC and PsychoNet 1 in all moods. The next section shows a more detailed discussion of the results.

Table 2: Snapshot of the result showing the previous and new annotations.

Concept	Previous Annotations	New Annotations
a scream of freedom	Affective,Hearing,Perceptual, Neg.Emotion,Pos.Emotion	Affective
coffee shop	Ingestion,Biological,Leisure,Money	Leisure
swimming pool	Relativity,Motion,Leisure	Leisure
The best way to commit a crime	Quantifier,Affective,Cognitive,Anger, Certainty,Achievement,Relativity, Neg.Emotion,Pos.Emotion	Relativity,Cognitive, Certainty,Affective
hit balls	Nil	Leisure
zip code	Nil	Relativity,Space

Table 3: Mood classification results using F-Measure.

Mood	LIWC	PNet 1	PNet 2
amused	0.40	0.56	**0.59**
cheerful	0.39	0.48	**0.49**
busy	0.40	0.56	**0.67**
happy	0.42	0.56	**0.61**
calm	0.33	0.44	**0.48**
content	0.28	0.42	**0.52**
creative	0.36	0.24	**0.41***
bored	0.39	0.50	**0.53**
contemplative	0.30	0.45	**0.58**
exhausted	0.44	0.30	**0.48***

4.1 Discussion

LIWC has been used successfully in various classification/identification tasks where the target classes are objective facts, such as Gender, Age, or Authorship Identification. However, the results of using LIWC in mood classification are poor and not promising as depicted in Table 3. This is mainly because the target class (mood) is subjective rather than objective, and may not be accurately provided by the user. It is usual that a user tags a number of posts with different moods even where the contents are, to some extent, similar. Hence, this task is challenging and LIWC features alone are not enough to fulfill it. 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).

As demonstrated in the experiment above, using PsychoNet 2 improved the result of mood classification compared to both LIWC and PsychoNet 1. PsychoNet 1 enhanced the result for some moods and improved accuracy to above 50% for others. However, PsychoNet 2 made enhancement in all moods and improved the accuracy to over 60%, for some moods. Furthermore, we can see that LIWC outperformed PsychoNet 1 in some moods (annotated with stars in Table 3). But the results confirm that PsychoNet-2 outperformed LIWC in all moods.

5 CONCLUSIONS

In this paper, we presented PsychoNet 2, a substantially contextualized and enriched psycholinguistic commonsense ontology. The overall main contribution is the creation of one cohesive semantic network and ontology based on the integration of three important text analysis resources namely: ConceptNet, LIWC, and WordNet. This addresses various limitations of PsychoNet 1, including contextualization and missing concepts. The first contribution, in this paper, is the contextual annotation of nodes. This contextualization annotates each node with the most representative contextual psycholinguistic categories. The second contribution is the enrichment of the LIWC, through utilizing the WordNet. This enrichment process added 7772 new words to the LIWC lexicon and associated them with the relevant psycholinguistic categories. Consequently, this enrichment led to the enrichment of the PsychoNet 2 by additional 8663 concepts, which were missing in PsychoNet 1, and improved the annotation of another 56615 concepts. PsychoNet 2 can be used in many applications in text engineering. We present here one application in mood classification. The results confirm the validity of PsychoNet 2 and showed the improvements experienced in all moods compared to LIWC and PsychoNet 1.

REFERENCES

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.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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)*.

Mohtasseb, H. and Ahmed, A. (2010a). *The Affects of Demographics Differentiations on Authorship Identification*, pages 409–417. Springer.

Mohtasseb, H. and Ahmed, A. (2010b). Psychonet: a psycholinguistc commonsense ontology. In *The International Conference on Knowledge Engineering and Ontology Development KEOD*, pages 159–164.

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*.