Identifying Writers' Background by Comparing Personal Sense Thesauri

Polina Panicheva[1,2], John Cardiff[1], and Paolo Rosso[2]

[1] Social Media Research Group, Institute of Technology Tallaght, Dublin, Ireland
{Polina.Panicheva,John.Cardiff}@ittdublin.ie
[2] Natural Language Engineering Lab, ELiRF, Universidad Politécnica de Valencia, Spain
prosso@dsic.upv.es

Abstract. Analysis of blogpost writings is an important and growing research area. Both objective and subjective characteristics of a writer are detected. Words have word meaning that is common in the language and that is represented in their usage. Another component of word meaning, "personal sense", not inherent in the language, but different for each person, reflects a meaning of words in terms of unique personal experience and carries the personal characteristics.

In our research word meaning techniques are applied to represent personal sense of words in texts by different authors. Personalized concept structures are construed and used to infer authors' perspective from text: various notions of context combined with different thesaurus similarity scales are applied to confirm that from a certain perspective similarity in the personalized thesauri with some restrictions can correspond to similarities in the occupation of the authors.

Keywords: personal sense, semantic relatedness, co-occurrence based thesauri, writer's perspective.

1 Introduction

Research on analysis of blogpost writings is an area attracting an increasing amount of attention. Not only can objective characteristics of a writer, like age or gender, be detected, but also subjective features, such as opinion, personality traits, and emotions, can be inferred. Words have word meaning that is common in the language and that is represented in their usage. We believe that another component of word meaning – "personal sense" [4] – not inherent in the language, but which is different for each person, can carry this personal information. This component reflects a meaning of the word in terms of unique experience of a person, in addition to word meaning belonging to language.

In our research the co-occurrence distributional techniques are applied to represent personal sense of words in texts by different authors, investigating four notions of 'context'. Personal concept structures, construed of the words taking into account their personal sense, are used to infer an author's perspective from the texts they write. The results confirm that similarity in the personalized thesauri can correspond to similarities in occupation of the authors.

C.J. Hopfe et al. (Eds.): NLDB 2010, LNCS 6177, pp. 288–295, 2010.

The rest of the paper is organized as follows. In Section 2 the work influencing our research is discussed. Section 3 describes some previous work using the personal sense technique. In Section 4 we present the ideas underlying our work based on the notion of personal sense. In Section 5 the experiments are described in detail; the results are presented and discussed in Section 6. Additional experiments of use to support and supplement the current work are discussed in Section 7.

2 Related Work

2.1 Author's Perspective

The importance of the distinction between what belongs to the author, the reader and the text itself is underlined in [1]. In this work the authors describe the important differences in considering subjectivity in text from these points of view. The emphasis is put on the text itself and the importance of the investigation of the opinion contained in it exclusively, as opposed to observing the reader's and writer's viewpoint in the text. Our work is concerned with personal information about the author, how and why it is presented in text.

In [2] the authors present an approach to perspective determination based on concept interrelation ontology. They manually construct an ontology of the movie-production field. The semantic relatedness measure for 25 pairs of words denoting motion picture industry concepts is presented. The next step suggested is gauging the author's perspective within the motion picture industry field.

The authors of [14] describe a method for constructing a personalized thesaurus from bookmarked web pages and documents. They construct a personalized thesaurus for every user as a context distributional profile of a word and propose a scale for measuring the difference between the thesauri.

2.2 Distributional Hypothesis

The distributional hypothesis is a widely used idea in semantics. The definition of word meaning, stating that "for a *large* class of cases… the meaning of a word is its use in the language" [13], appears to be helpful for analyzing word meaning in natural language applications. Intuitively it has been used in different linguistic tasks, e.g. [11], [5]. The degree of applicability of this hypothesis to natural language is investigated in [6], testing different algorithms realizing this idea and making a crucial distinction between its applications to words and to concepts.

3 Previous Work on Personal Sense

Our initial experiments using personal sense were described in [7]. There we investigate the personal sense of the words 'movie' and 'film', and the hypothesis is that the personal sense of these words is different for the authors writing a negative and a positive review. The resulting features are reported to perform higher than the baseline, but very modestly compared to the widely used n-gram word-count features. The authors assume that the reason for this is that different writers express different personal sense for the movie-words, even if the polarity of their reviews is the same.

An experiment testing this assumption is described in [8]. The result showed that opinion mining with a bunch of different writers yielded so much noise, that a much smaller number of instances but analyzed separately for different writers gave better performance. It was concluded that polarity in texts may be expressed in an individual way by different authors.

4 Author Characterization Using Personal Sense

Personal background and other personal information about the author is revealed in text, forming a person's idiolect – "a language that can be characterized exhaustively in terms of … properties of some single person at a time" [4]. An idiolect represents a collection of personal characteristics at the same time (i.e., age, gender, social class, occupation), as well as personal traits and private states. Words have word meaning that is common in the language and that is represented in their usage. 'Personal sense', a component of word meaning not inherent in the language, but different for each person, is defined in [4]. This component reflects word meaning in terms of unique experience of a person.

4.1 Personal Sense and Co-occurrence Distribution

Our approach is based on the hypothesis that personal sense is represented in the way that words are put together to form a text and in the way they occur together with each other. We use the co-occurrence distributional method described in [6] to represent personal senses.

We investigate different notions of 'context' in order to find the one that makes the resulting co-occurrence distribution representing the meaning or personal sense of a word best: from ordered pairs of words, one word on the left and one on the right, to all the words occurring in the same sentence with the word in question.

Authors of [6] find out that for investigating semantic measures of word-to-word distance the best results are obtained by using the cosine distance measure using conditional probability weighting of coordinates. This is the measure that we use in our work to scale semantic relatedness of personal sense of words to each other.

4.2 From Personal Sense to Author's Perspective: Forming a Structure

We follow [2] in investigating the semantic inter-relatedness of words for different persons, especially for authors with different occupations. In a text by one person, the concepts that he/she uses acquire a personal sense. Depending on unique background information underlying their idiolect, the concepts represented in the text will form a unique semantic structure in terms of their meaning, as influenced by the personal sense. Whereas the authors of [2] define every profession in terms of its semantic relatedness to every other notion, suggesting a manually created ontology of the semantic field, our consideration is to derive the semantic relatedness of the notions from text, and thus infer the author's profession.

Thus, the goal of this work is as follows. Firstly, to represent personal sense of words in text by different authors using co-occurrence distributional word meaning techniques. Secondly, to form personalized thesauri based on the inter-relations between words. Thirdly, to infer relations among authors' in terms of their

background, particularly their occupation, from the distance measures between their personalized thesauri. Thus, we continue exploiting the interrelation of private states expressed in text and authorship attribution, described in [8].

5 Experiments

We work with blog text corpus described in [12]. For the initial experiment we investigated 10 randomly-selected blogs annotated as 'Accounting' and 10 annotated as 'Military', with the annotation referring to the author's self-annotation in terms of their occupation or profession. We used a Natural Language Toolkit (NLTK)[1] for the Python programming language for most of the text analysis. All the texts were annotated automatically with parts-of-speech using the Treebank POS-Tagger built-in into the NLTK.

A group of nouns was selected for the experiment, consisting of 33 nouns, including the 25 most frequent nouns ([*dad, mom, person, company, movie, husband, test, parent, dinner, world, head, place, child, thing, school, way, life, family, house, work, job, car, home, girl, friend*]) and the 8 manually selected nouns ([*career*, *power*, *war*, *law*, *rule*, *army*, *man*, *woman*]).

Next, their personal sense was represented by the words' co-occurrence distributional vectors. Every dimension is a word occurring in the context(s) of the investigated nouns. As an example, the contexts of the target word 'woman' taken from the following sentences are presented:

-"Part of the problem for female soldiers in the Army is the existence of *women* like one of my roommates, who…"

-"SGT B, however, the *woman* I live with, has clearly not had many female friends."

We observed 4 different definitions of context:

1. Context of a word *w1* is a set of pairs of words; every pair consists of a word immediately to the left of *w1* and a word immediately to the right of *w1*:
   - *(of/IN, like/IN), (the/DT, I/PR)*
2. A set of the same words as in (1), not organized in pairs, but for each word it is indicated if it belongs to the left or to the right context.
   - *l_of/IN, r_like/IN, l_the/DT, r_I/PR*
3. The same set of words as in (2), but no left- or right-side context indication.
   - *of/IN, like/IN, the/DT, I/PR*
4. All the words belonging to the sentences in which *w1* occurs. except *w1* itself.
   - *Part/NN of/IN the/DT problem/NN for/IN female/JJ soldiers/NN in/IN the/DT Army/NN is/VB the/DT existence/NN of/IN* etc.
   - *SGT/NN B/NN ,/, however/RB ,/, the/DT, I/PR live/VB with/IN ,/, has/VB clearly/RB not/RB had/VB many/JJ female/JJ friends/NN*

Every word in context acquires a weight of the conditional probability of the word given the target word *w1* (see formula 1).

$$P(w \mid w_1) = \frac{P(w \cap w_1)}{P(w_1)} = \frac{Frequency\,(w, w_1)co-occurring}{Frequency\,(w_1)} \quad (1)$$

[1] Available at `http://www.nltk.org/`

For computing similarity between two words in a text we used the cosine similarity measure between the target words (see the Cosine for Conditional Probabilities formula, (2)).

$$CosCP(w_1, w_2) = \frac{\sum_{w \in cont(w_1) \cup cont(w_2)} (P(w \mid w_1) * P(w \mid w_2))}{\sqrt{\sum_{w \in cont(w_1)} P(w \mid w_1)^2} * \sqrt{\sum_{w \in cont(w_2)} P(w \mid w_2)^2}} \quad (2)$$

The experiments are dedicated to comparing the obtained personalized thesauri with the techniques similar to those described in [14]. We compared three different thesauri distance measures, using the basic formula presented in [14].

$$d = \sqrt{\frac{\sum_{i=1}^{m} \sum_{j=1}^{m} d'^{2}_{v_i w_i}}{q}} \quad (3)$$

$$d_{vw}' = \begin{cases} \left| sim_{vw}^{S} - sim_{vw}^{T} \right|, (v,w) \in S,T \\ sim_{vw}^{S}, (v,w) \in S, \notin T \\ sim_{vw}^{S}, (v,w) \in T, \notin S \\ x, (v,w) \notin S, \notin T \end{cases} \quad (4)$$

The first distance measure, d_1, was exactly the same as described in [14]. That means, in equation (3) q is equal to m^2, the squared number of words in the target words vocabulary; and x in equation (4) equals to 1. For the second measure, d_2, x in equation (4) is equal to 0. In other words, when the word-pair (v, w) is not present in neither of the two thesauri S and T, we add nothing to the sum. For d_3, x is also equal to 0. But q in equation (3) represents the number of the word pairs present in at least one of the two vocabularies S and T. The difference between the thesauri distance formulas 1, 2 and 3 lies in approaching the words absent from the personal blog vocabulary in a different way. We assume that this difference will affect the results in a considerable way, because the two groups of blog authors that we are investigating manifest differences in terms of their vocabulary usage: the 'Military' authors tend not to use all of the target words in their blogs, whereas for the 'Accounting' authors this is usually not the case.

We assume that at least one notion of 'context' constitutes the personal senses of the observed words so that the personal sense and their inter-relations represent the authors' background to a sufficient degree, and the resulting average distances between personalized thesauri are bigger for authors having different occupation than for those sharing one.

6 Results and Discussion

The personalized thesauri comparison results for the 33 target words, for 4 different types of context and the 3 distance functions are presented in Table 1.

Our assumption was that the personal word sense interrelation structures would be more similar among authors belonging to the same occupational background, than between the authors belonging to different professional groups. We represented the personal sense interrelations of every author as a personalized thesauri based on their texts. The distances between the thesauri were computed for every pair of authors. We divided the pairs of thesauri into 3 types depending on their authors:

- (1) (T[military], T[military])
- (2) (T[accounting], T[accounting])
- (3) (T[military], T[accounting]), and (T[accounting], T[military]),

where T[*x*] represents the thesaurus of an author belonging to the occupation *x*.

We computed the distances between the thesauri in groups 1, 2 and 3 separately. According to our assumption, the distances in groups (1) and (2), i.e. between thesauri by authors of the same occupation, would be smaller than for the distances of group (3), representing the distances between the different occupational group thesauri. In spite of the fact that the results did not unanimously confirm our overall hypothesis, they indicate some very important tendencies and provide evidence supporting our contention that the personal sense technique is useful for discriminating between authors' occupations.

The results with high statistical significance (p-value < 0.01), obtained for the context definition 1, for distance measures 2 and 3, for the 'Military' occupation group, confirm the hypothesis and show that the inter-group distances are lower than the intra-group distances. This could also indicate the fact that the military group is easier to discriminate than the accounting group, as the personalized thesauri for these group are closely related to each other, whereas for the accounting group the structure is sparser. In most of the cases little statistical significance was achieved, indicating that more target words and more texts should be analyzed.

It is important to note that we have obtained contradictory results for different distance options, according to our expectations described in Section 5. In most of the cases distance 1 yields exactly opposite results than distances 2 and 3, in terms of the inter- and intra- group distances between the occupations. These formulas mainly differ in the way that they handle words that did not occur in the texts by specific authors. As the results for the different distance formulas suggest, the difference in approaching these words alters the results dramatically, regardless all the other options. This means that for the current data and context definition, the absence or presence of words in the authors' vocabularies outperform the personal sense differences computed for the words. This was our expectation that was confirmed by using the words that were not contained in all of the texts.

The only context definition that yielded more consistent results for different distance functions is the context number 4, containing all the words in the sentence where the target word appears. This confirms a consideration that the broader the context definition is, the more it influences the results, regardless of the different distance measures. However, the very low statistical significance in the case of such context shows that the number of target words should be increased.

Table 1. Personalized thesauri comparison results for blogs by authors of the same and different occupation

	Thesauri distance comparison		Statistical significance according to the paired t-test(p-value)	
d_1				
Context:	'Military' vs different	'Accounting' vs different	'Military' vs different	'Accounting' vs different
1	>	<	0.74	0.2
2	>	<	0.2	0.1
3	>	<	0.1	0.1
4	>	<	0.47	0.68
d_2				
1	<	>	*<0.001*	0.1
2	<	>	0.09	0.8
3	<	>	0.35	0.8
4	>	>	0.61	0.59
d_3				
1	<	>	*<0.001*	*0.02*
2	<	>	0.15	0.99
3	<	<	0.56	0.99
4	>	>	0.5	0.7

7 Conclusions and Future Work

Our experiments confirm that for target words of certain number and frequency, and for a certain definition of context and a certain distance measure for thesauri comparison, the differences between the resulting thesauri can effectively represent the differences among the occupation background of the authors of texts. Personalized concept structures were constructed from texts by different authors, taking into account the personal sense of the words they used. The resulting structures were unique for every blog writer, representing the unique personal sense. The personal concept structures were used to infer the authors' perspective from their writings. The results confirmed that with certain restrictions, the method of representing the similarities in the personalized thesauri could be used to reflect similarities in occupation of the authors.

The first restriction applies to the occupation groups themselves: the results prove that different occupations are not equally easy to analyze, i.e., some of the occupations present similarities in the personalized thesauri that can be easily detected, whereas others require additional experiments with more refined options.

We see the direction of such experiments as follows. First of all, it is necessary to reduce the influence of the fact that some words are absent in some blogs, by including in the target words list only the words that appear in all of the investigated blog texts. Secondly, it is important to increase the influence of the personal sense method, i.e. of the context information, on the resulting personalized thesauri, by considering context in a broader sense and increasing the context window. However,

it has been illustrated that the broader context window leads to having more noisy information, and subsequently less statistical significance. This issue will be addressed by analyzing more target words and more texts, thus acquiring higher occurrence values.

Acknowledgements

The work of the third author has been in the context of the TEXT-ENTERPRISE 2.0 (TIN2009-13391-C04-03) research project.

References

1. Balahur, A., Steinberger, R.: Rethinking sentiment analysis in the news: from theory to practice and back. In: Proc. 1st Workshop on Opinion Mining and Sentiment Analysis (WOMSA), CAEPIA-TTIA Conference (2009)
2. Choudhury, S., Raymond, K., Higgs, P.L.: A rule-based metric for calculating semantic relatedness score for the motion picture industry. In: Workshop on Natural Language Processing and Ontology Engineering (2008)
3. Firth, J.R.: A synopsis of linguistic theory 1930-1955. In Studies in Linguistic Analysis. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.), Selected Papers of J.R. Firth 1952-1959, London: Longman (1968)
4. Leontev, A.N.: Activity, Consciousness, and Personality. Prentice-Hall, Hillsdale (1978)
5. Mitrofanova, O., Lashevskaya, O., Panicheva, P.: Statistical Word Sense Disambiguation in Contexts for Russian Nouns Denoting Physical Objects. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2008. LNCS (LNAI), vol. 5246, pp. 153–159. Springer, Heidelberg (2008)
6. Mohammad, S., Hirst, G.: Distributional measures of concept-distance: A task-oriented evaluation. In: Proc. EMNLP 2006 (2006)
7. Panicheva, P., Cardiff, J., Rosso, P.: A Co-occurrence Based Personal Sense Approach to Opinion Mining. In: Proc. 1st Workshop on Opinion Mining and Sentiment Analysis (WOMSA), CAEPIA-TTIA Conference (2009)
8. Panicheva, P., Cardiff, J., Rosso, P.: Personal Sense and Idiolect: Combining Authorship Attribution and Opinion Analysis. Manuscript submitted for publication (2010)
9. Pinto, D., Rosso, P., Juan, A., Jiménez, H.: A Comparative study of Clustering algorithms on Narrow-Domain abstracts. In: Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) 37 (2006)
10. Romesburg, H.C.: Clustering Analysis for researchers. Wadsworth, Inc. (1984)
11. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Communications of the ACM 8(10) (1965)
12. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of Age and Gender on Blogging. In: Proc. 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs (2006)
13. Wittgenstein, L.: Philosophical Investigations: The English Text of the 3rd Edition, translated by G.E.M. Anscombe. Macmillan Publishing, New York (1973)
14. Yoshida, S., Yukawa, T., Kuwabara, K.: Constructing and examining personalized cooccurrence-based thesauri on Web pages. In: Proc. 12th International World Wide Web Conference (2003)