Investigating Verbal Intelligence using the TF-IDF Approach

Kseniya Zablotskaya[1], Fernando Fernández Martínez[2]*, Wolfgang Minker[3]

[1,3] Institute of Communications Engineering, University of Ulm, Germany
[2] E.T.S.I. de Telecomunicacíon, Universidad Politécnica de Madrid, Spain
[1,3]{kseniya.zablotskaya,wolfgang.minker}@uni-ulm.de, [2]ffm@die.upm.es

Abstract

In this paper we investigated differences in language use of speakers yielding different verbal intelligence when they describe the same event. The work is based on a corpus containing descriptions of a short film and verbal intelligence scores of the speakers. For analyzing the monologues and the film transcript, the number of reused words, lemmas, n-grams, cosine similarity and other features were calculated and compared to each other for different verbal intelligence groups. The results showed that the similarity of monologues of higher verbal intelligence speakers was greater than of lower and average verbal intelligence participants. A possible explanation of this phenomenon is that candidates yielding higher verbal intelligence have a better short-term memory. In this paper we also checked a hypothesis that differences in vocabulary of speakers yielding different verbal intelligence are sufficient enough for good classification results. For proving this hypothesis, the Nearest Neighbor classifier was trained using TF-IDF vocabulary measures. The maximum achieved accuracy was 92.86%.

Keywords: Verbal intelligence, cosine similarity, TF-IDF measures

1. Introduction

We all are different and, even describing the same event, we use different words and sentence structures. Our vocabulary depends on our education, social status, age, gender, life experience, etc. The goal of our research is to find out which language peculiarities may reflect the verbal intelligence of speakers. When two persons are trying to repeat a story, along with their own words and phrases they may recall several expressions from the original text. On the one hand the number of repeated words shows speakers' short-term memory and a good ability to "convey" information to a listener, on the other hand proper synonyms reflect the richness of speakers' vocabulary and their skills to use language for expressing own thoughts and feelings. The ability to use language for accomplishing certain goals is called verbal intelligence (VI) (Goethals et al., 2004; Cianciolo and Sternberg, 2004). In other words, verbal intelligence is "the ability to analyse information and to solve problems using language-based reasoning" (Logsdon, 2012).

The goal of our research is to find out which language peculiarities may reflect the verbal intelligence of speakers. The automatic estimation of users' verbal intelligence may help Spoken Language Dialogue Systems (SLDSs) more effectively control the flow of the dialogues, engage the users in an interaction, be more attentive to human needs and preferences and as a result be more helpful and user-friendly (Figure 1). For training machine learning algorithms we need to know a maximum number of language features that reflect speakers' verbal intelligence. In this work we investigate to which extent the vocabulary of test persons reflect their levels of verbal intelligence when they all describe the same event and explain their thoughts and feelings about it.

*For this work Fernando was granted a fellowship by Cajamadrid Foundation

2. Corpus Description

For the data acquisition a short film was shown to German native speakers. It described an experiment on how long people could stay without sleep. The test persons were asked to imagine that they met an old friend and wanted to tell him about this film. Our goal was to record every-day speech when talking to relatives and friends.

The test persons were also asked to participate in the verbal part of the Hamburg Wechsler Intelligence Test for Adults (HAWIE) (Wechsler, 1982). The verbal part consists of the following sub-tests:

- *Information*: this sub-test measures general knowledge and includes questions about history, geography, literatures, etc;
- *Comprehension*: test persons are asked to solve different practical problems and explain some social situations;
- *Digit Span*: test persons are asked to repeat increasingly longer strings of numbers first forward and then backward; the sub-test measures short-term memory;
- *Arithmetic*: test persons are asked to solve some arithmetic problems given in a story-telling way; the sub-test measures their concentration and computational ability;
- *Similarities*: test persons are asked to find a similarity between a pair of words;
- *Vocabulary*: test persons are asked to explain increasingly more difficult words using their own vocabulary.

The raw scores of each test person on the verbal test are based on his correct answers (Figure 2). The raw scores are then converted into "Scaled Scores" using special tables (Wechsler, 1982). The Scaled Scores vary between

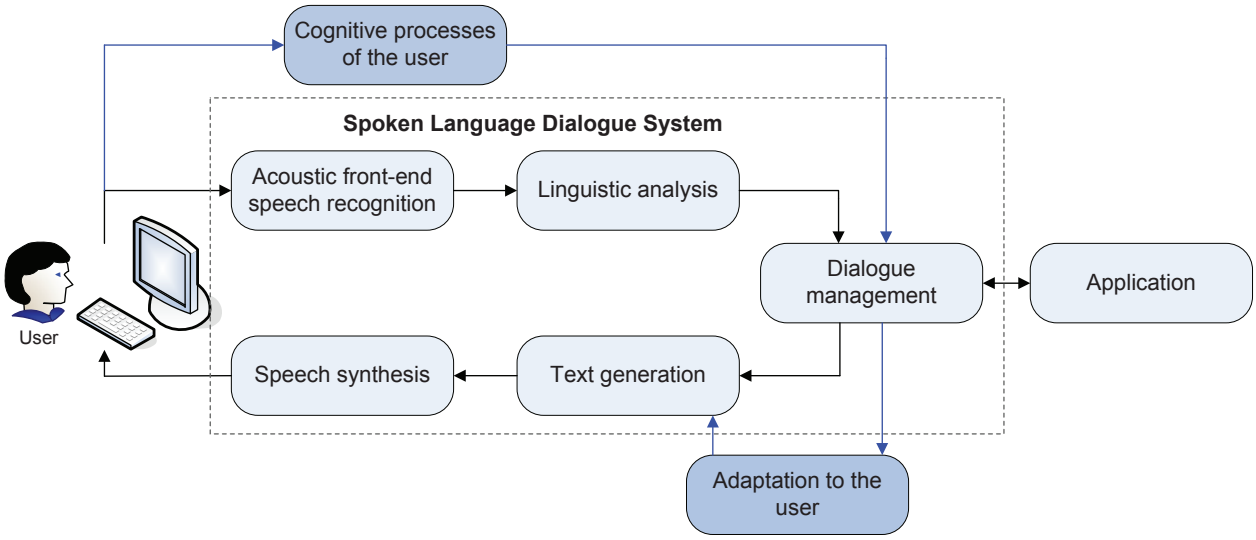

Figure 1: Spoken Language Dialogue System

0 (lowest scaled score) and 16 (highest scaled score) and may be used to compare the performance of the participants. The sum of the scaled scores and the age of a test person are used to estimate his verbal intelligence score.

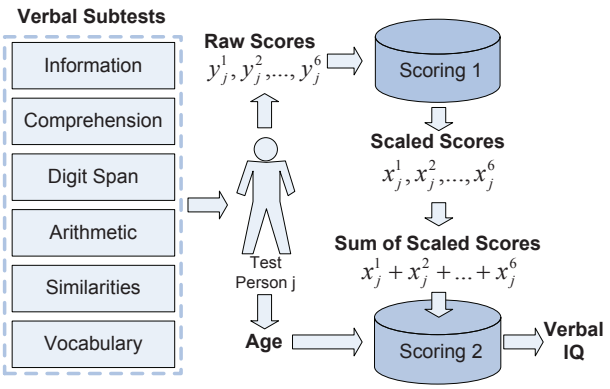

Figure 2: Verbal Part of the Hamburg Wechsler Intelligence Test for Adult

This corpus described in (Zablotskaya et al., 2010) consisted of 56 monologues and 30 dialogues (10 hours of audio data). During the experiment the test persons were also asked to engage a conversation with a dialogue partner. However, in this investigation only the monologues were analysed. For this work we have enlarged our corpus, which now contains 100 monologues (6 hours), 56 dialogues (12 hours) and verbal intelligence scores of all the test persons.

3. Feature Extraction

To analyse the vocabulary of people yielding different verbal intelligence when describing the same event, we compared the monologues with the film transcription. Figure 3 shows excerpts from the film and from one of the monologues[1].

Excerpt from the film
Max and Funda have been without sleep for fifty eight hours. They have laid down on the sofa. Is it a mistake? Actually they would like to move. But now they cannot any more. The blood pressure is down, the energy reserves are over. They both are freezing despite the fire-place and the jacket. The question is who closes the eyes first. It is Max. Funda wins. She stays awake a few minutes longer.
Excerpt from a corresponding monologue
After fifty eight hours, they were really tired. And, they had frozen. Despite they had very warm clothes. And then the man fell asleep and then the woman.

Figure 3: Excerpts from the film and one of the recorded monologues.

For the comparison, the following features were extracted:

- *Number of reused words* - number of words that a test person "reused" from the film. For our example in Figure 3 the reused words are: *fifty, eight, hours, they, and, they, despite, they, and, the, and, the*.
- *Number of unique reused words*. It includes the number of reused words without repetitions. In Figure 3, the unique reused words are *fifty, eight, hours, they, and, despite, the*.
- *Number of all reused lemmas*. This feature has been calculated as *Number of all reused words* with the difference that lemmas were considered.
- *Number of unique reused lemmas*. This feature has been calculated as *Number of unique reused words* with the difference that unique lemmas were taken into account.

[1]As the conversation language is German, the example was directly translated into English.

- *Cosine similarity* between the film and a k_{th} monologue using lemmas. For this feature extraction, we have created a matrix consisting of all unique lemmas from the film, including the frequency of these lemmas within the film and within a k_{th} monologue. The frequencies were normalized by the total amount of words in the corresponding text; the cosine similarity between the two normalized vectors (lemma frequencies within the film and lemma frequencies within a k_{th} monologue) was calculated as:

$$similarity = \frac{\sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2},$$

where n is the number of unique lemmas in the film, a_i - frequency of i_{th} lemma in the film, b_i - frequency of i_{th} lemma in the monologue.

- *Number of reused n-grams*. For this feature we have calculated the number of n-grams ($n = \overline{2,10}$) that were used in the film and then reused by a test-person in his monologue. In our example, *the number of reused 2-grams* equals to 2 (reused 2-grams are *fifty eight* and *eight hour*), *the number of reused 3-grams* equals to 1 (*fifty eight hour*), etc.

- *Cosine similarity using n-grams*. The cosine similarity was calculated from a feature vector composed by the counts of different n-grams for each monologue.

- We have also determined the number of lemmas that were used by the candidates but were not used in the film. For each monologue the following features have been calculated: *Own lem*$_1 = \sum_{i=1}^{n} frequency(lem_i) * count(lem_i)$ and *Own lem*$_2 = \sum_{i=1}^{n} frequency(lem_i)$, where n is the number of unique lemmas that were used by a test person but were not used in the film; $count(lem_i)$ shows how many times $lemma_i$ was used in the monologue; $frequency(lem_i)$ shows the frequency of $lemma_i$ according to a frequency dictionary of the German language (Kupietz et al., 2010). This dictionary consists of 40000 German words with frequency from 1 to 17: 1 corresponds to more frequent words, 17 corresponds to less frequent words. If a word from the monologues was not found in the dictionary, its frequency was set to 20.

4. Feature Analysis

The k-means algorithm was applied on the scaled scores of the test persons (Figure 4). For the feature analysis two experiments were performed. In the first experiment the observations were partitioned into 2 clusters: Cluster P_1 consisted of test persons with lower verbal intelligence, P_2 contained candidates with higher verbal intelligence. In the second experiment the test persons were partitioned into 3 clusters: P_1 - lower verbal intelligence, P_2 - average verbal intelligence, P_3 - higher verbal intelligence.
The mean values of all the features from the clusters were compared using ANOVA (Sachs and Hedderich, 2006) (Figure 5). In Experiment I with two clusters, features with small p-values were:

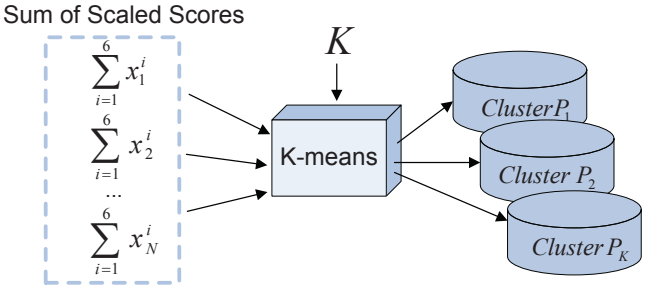

Figure 4: The K-means algorithm

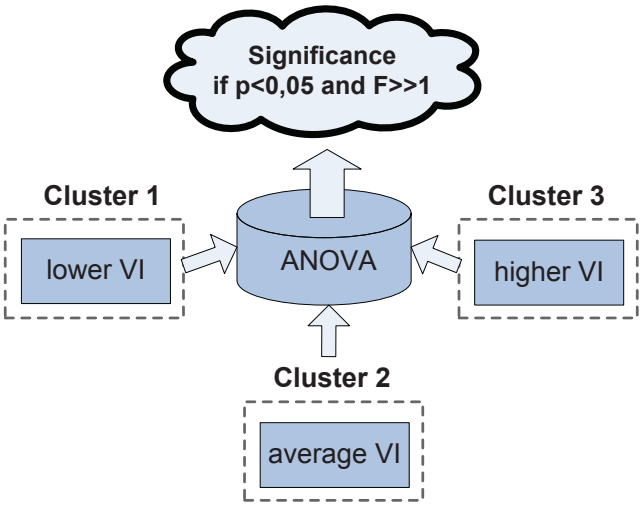

Figure 5: ANOVA for Experiment 2

- *Number of reused 3-grams* (averaged value for the first class $AV_{low} = 0.021$, averaged value for the second class $AV_{high} = 0.031$, $p = 0.012$, $F = 6,63$);

- *Cosine similarity using lemmas* ($AV_{low} = 0.79$, $AV_{high} = 0.83$, $p = 0.03$, $F = 4,64$);

- *Cosine similarity using repeated n-grams* ($AV_{low} = 0.13$, $AV_{high} = 0.15$, $p = 0.01$, $F = 7,07$).

In Experiment II with three clusters, a feature with a small p-value was:

- *Cosine similarity using repeated n-grams* ($AV_{low} = 0.13$, $AV_{aver} = 0.14$, $AV_{high} = 0.16$, $p = 0.01$, $F = 7,07$).

As we can see, people with higher verbal intelligence used more words from the film and the similarity between their descriptions and the film is higher than the similarity of people with average and lower verbal intelligence. This may be explained in the following way. One of the verbal sub-tests of HAWIE is Memory. A high memory score relates to a high verbal intelligence score of a test person. Also, people with good memory were easier able to remember many details of the film and to use words which they heard when watching the program. We may conclude that vocabulary of people yielding different verbal intelligence is different when they talk about the same event even taking into account that they were asked to talk about this film just after they had watched it.

5. Classification Results and Discussion

In this work we also investigate another hypothesis: test persons belonging to different verbal intelligence classes may be distinguished by word or lemma patterns regardless of the order of these words and lemmas in the monologues. In other words, differences in vocabulary of people yielding different verbal intelligence are sufficient enough for providing good classification results. For proving or rejecting this hypothesis, the Nearest Neighbour classifier was trained for the automatic classification of monologues into three groups: test persons with lower, average and higher verbal intelligence. For the classification task each monologue was represented as a list of words. Each word was considered as a feature and each monologue was represented as a feature vector. The value of each feature was equal to a weight of the corresponding word which was calculated using the TF-IDF (Term Frequency-Inverse Document Frequency) weighting scheme (Manning et al., 2008):

$$w_{ij} = tf_{ij} \cdot idf_i,$$

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}, \quad idf_i = log\frac{|D|}{\{j : t_i \in d_i\}}$$

where n_{ij} - frequency of a term t_i in a document d_j, $\sum_k n_{kj}$ - the number of words in a document d_j, $|D|$ - total number of texts in the corpus, $\{j : t_i \in d_i\}$ - number of documents in which t_i appears. The weights w_{ij} show the importance of the words in each text-file. As can be seen, more frequent terms in a document are more representative and if the number of documents in which this term occurs increases, this term becomes less discriminative.

To reduce the number of features, we used lemmas instead of words. For the feature selection, the lemmas with the highest TF-IDF values were used. All the lemmas were sorted according to their TF-IDF measures and then the top N most indicative terms were selected. The remaining lemmas were removed as stop words or common words that did not add any meaningful content. This procedure was performed separately for each class. By observing the evolution of the classification accuracy with an increasing N-value, we determined the minimum vocabulary size required to achieve the optimum performance. For the classification the Leave-One-out cross validation method was used. The performance of the k-nearest neighbours algorithm had a maximum accuracy of 92,86% for the dimensionality of 155.

This work has shown that verbal intelligence can be recognized through language cues. The achieved classification accuracy can be deemed as satisfying for a number of classes that is reasonably high enough to enable its integration into a SLDS. Unlike typical text categorisation, our verbal intelligence prediction task is influenced by the necessary fact that the different categories or classes to be identified are not well separated from a conceptualization point of view. Of course, it would have been easier to distinguish people talking about different topics from their every-day life although the results for such a comparison across different topics would not have been objective. By letting the participants (i.e. people with different interests and hobbies) to discuss their own topics, we would be then recognizing the topics themselves rather than people with different cognitive processes. On the other hand, the use of German, a very agglutinative language, has resulted to be a drawback with regards to word lemmatization. By lemmatization of compound words (compounding is a pretty common phenomena in German) we are basically loosing the extra meaning that arises from the combination of the interrelated words. This meaning has proven to be really helpful to correctly discriminate between different levels of verbal intelligence. This also suggests the importance of finding some other features that could be more robust when used in a conventional system. Prosodic features could be a good alternative so it would be interesting to start working on an multi-modal inference framework that could jointly exploit the potential of, among others, this kind of features. As we have already mentioned, the linguistic cues that we have used in this work could pose a problem, for instance, if we want to apply these solutions with the same users but across different domains. In this regard, prosodic features would be found to be advantageous as they would also allow us to explore the possibility to find topic independent solutions.

6. Acknowledgment

This work is partly supported by the DAAD (German Academic Exchange Service).

Parts of the research described in this article are supported by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

The work leading to these results has been partially supported by TIMPANO (TIN2011-28169-C05-03) project.

7. References

A.T. Cianciolo and T.J. Sternberg. 2004. *Intelligence: a Brief History*. Blackwell Publishing.

G.R. Goethals, G.J. Sorenson, and J.M. Bruns (Editors). 2004. *Encyclopedia of Leadership*. Sage Publications (CA).

M. Kupietz, C. Belica, H. Keibe, and A. Witt. 2010. The german reference corpus DeReKo: A primordial sample for linguistic research in: Calzolari, nicoletta et al. (eds.). In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010)*, pages 1848–1854.

A. Logsdon. 2012. *Learning Disabilities*. http://www.learningdisabilities.about.com/.

C.D. Manning, P. Raghavan, and H. Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

L. Sachs and J. Hedderich. 2006. *Angewandte Statistik*. Springer; Berlin, Heidelberg.

D. Wechsler. 1982. *Handanweisung zum Hamburg-Wechsler-Intelligenztest fuer Erwachsene (HAWIE)*. Separatdr., Bern; Stuttgart; Wien, Huber.

K. Zablotskaya, S. Walter, and W. Minker. 2010. Speech data corpus for verbal intelligence estimation. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, May.