С. Ультес, А. Шмитт, Р. Эль Хаб, В. Минкер

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ КАЧЕСТВА ВЗАИМОДЕЙСТВИЯ В РЕЧЕВЫХ ДИАЛОГОВЫХ СИСТЕМАХ: СРАВНЕНИЕ КЛАССИФИКАТОРОВ, ОСНОВАННЫХ НА (УСЛОВНЫХ) СКРЫТЫХ МАРКОВСКИХ МОДЕЛЯХ, И МАШИН ОПОРНЫХ ВЕКТОРОВ

В последнее время была представлена метрика качества диалога для оценки качества взаимодействия с языковой диалоговой системой. Эта метрика позволяет получить оценку качества в произвольный момент взаимодействия человека с машиной. В то время как предыдущая работа базировалась на методе опорных векторов (SVM) для классификации качества взаимодействия на основе статического характеристического вектора, представляющего всю предысторию взаимодействия, здесь мы исследуем условные скрытые марковские модели (CHMM), которые принимают во внимание последовательный характер данных и, в отличие от стандартных скрытых марковских моделей (HMM), высчитывают вероятности классов. Экспериментальные результаты показали, что CHMM достигла значения невзвешенного среднего вызова (UAR) равного 0.39. Таким образом, алгоритм уступает HMM с UAR равным 0.44 и SVM с UAR равным 0.49. Все алгоритмы тренировались и исследовались в равных условиях.

Ключевые слова: качество взаимодействия, машины опорных векторов.

UDC 519.234

K. V. Zablotskaya, S. G. Zablotskiy, F. Fernández-Martínez, W. Minker

LANGUAGE STYLE MATCHING AND VERBAL INTELLIGENCE

In this paper language style matching of speakers yielding different verbal intelligence was analyzed. The work is based on a corpus consisting of 100 descriptions of a short film (monologues), 56 discussions about the same topic (dialogues) and verbal intelligence scores of the test persons. According to the results, higher verbal intelligent speakers showed a greater degree of language style matching when describing the film and were able to better adapt to their dialogue partners compared to lower verbal intelligent participants.

Keywords: spoken dialogue system, linguistic analysis, ANOVA .

Statistical approaches are often applied to text analysis and information retrieval problems. For example, TF-IDF measures may be used for classification of documents into a fixed number of predefined categories. Comparing a document with special dictionaries may be helpful for its content and semantic analysis. Linguistic analysis of texts allows researchers to determine additional information about authors: age, social status, emotions, psychological state, etc. In this research we applied a relatively new statistical method, tokens n-gram distributions, to the analysis of language style matching (LSM).

When two speakers are talking to each other, they try to adapt to their dialogue partner and to somehow synchronize their verbal behaviors. This phenomenon was investigated in [1]. In [2] linguistic style matching in human-human conversations was analyzed. For the linguistic analysis all the utterances were compared with a special dictionary which contained words sorted by a number of categories. The usage of each category was analyzed on conversation and turn-by-turn levels and showed that speakers synchronized their words when talking to each other. In [3] college students were asked to write answers to several questions formulated in different styles. It was shown that students followed the language styles of written questions.

In this paper we analyzed LSM of speakers with different verbal intelligence. This investigation may be helpful for improvement of user-friendliness of spoken language dialogue systems (SLDSs). SLDSs which automatically adapt to users' language styles and change their dialogue strategies may help users to feel more comfortable when interacting with them. However, it is necessary to know how different speakers change their own language styles in order to adapt to their dialogue partners. In [3] it was shown that students with higher grades matched the linguistic styles of asked questions closer than other students participated in the experiment. In this research we analyzed spoken utterances of people with different verbal intelligence. In the first part of this research we analyzed similarity between language styles of verbal descriptions of a short film (monologues) made by test persons who participated in our experiments (German

native speakers of different ages, social status and education levels) and the language style of the film transcript. In the second part, LSM of dyadic conversations were compared with verbal intelligence scores of the dialogue partners and their levels of acquaintance (whether the dialogue partners were relatives/close friends or strangers who had not met each other before the experiment). In other words, our goal was to investigate how adaptation in a conversation depends on the relationship between dialogue partners and whether their levels of verbal intelligence play the role in this process.

Corpus Description. For the corpus collection 100 German native speakers of different genders, ages, educational levels and social status were asked to participate in a study conducted at the University of Ulm, Germany. The participants were shown a short film from the TV-Program Galileo and were asked to imagine that they were talking to a good friend of theirs. They had to describe the main idea of the film with their own words. The candidates were not asked to somehow follow the language style of the film; they were asked to talk as naturally as possible in order to capture their every-day conversation styles. The chosen film was about an experiment on how long people could stay awake. Two men and one woman were asked to stay in the same house and to fight against sleep. When they were in a bathroom, they had to sing a song or to whistle. The participants also had to take different tests to control their concentration, memory, attention, condition and a general well-being. As a result the woman won. She could be without sleep for 58 hours. At the end of the film it was told that sleep was very necessary and experiments with animals showed that being without sleep can be dangerous to your life.

91 out of 100 participants were asked to make 10-minute conversations with another test person resulting in 55 two-person dialogues and 1 three-person dialogue. The topic of the dialogue was about the education and the school system in Germany. The participants had to express their opinions, to determine advantages and disadvantages of the school system, to talk about teachers, lectures, marks, etc. If the candidates hadn't met each other before and had difficulties in making a dialogue, they were asked to dispute and to prove a certain position about the school system. For example, they were asked to imagine that they had different points of view about German education. The first participant was asked to prove that the school system in Germany is very good, that the children get a very good education and it is no use making changes to it. The second participant was asked to describe bad features of German education, make different examples and to offer some innovations. Sometimes it helped the participants to dispute because they could analyze the position of the dialogue partner and to react in some way. But sometimes it was more difficult for the participants to keep the conversation going because they couldn't find (for example) good features of the education if their private opinion was different.

The other 9 test persons were not able to participate in dialogues because the experiment was time consuming for them. Also, several test persons participated in several dialogues with different dialogue partners.

Afterwards, verbal intelligence of the candidates was measured using the Hamburg Wechsler Intelligence Test for Adults [4]. It's the German version of the American original. Its scale is based on a projection of the subject's measured rank on the Gaussian bell curve with a center value (average IQ) of 100 and a standard deviation of 15. The test is organized for adults ranging in age from 16 to 74 years and consists of 6 verbal and 5 performance tests. Education, experience and life-style also contribute to scoring better on this test. For the research we used only the verbal section:

– *Information.* With this sub-test the general knowledge is measured; 25 questions come from a particular culture. For example, «What is the capital of Russia?»

– *Comprehension.* This sub-test measures social awareness and common-sense. It focuses on the social sense and the conception of cultural values. For example, «What would you do if you lost your way in a forest?»

– *Digit Span.* The auditory short memory, concentration and attention are measured with this sub-test. A participant is asked to repeat strings of digits forward and then backward.

Arithmetic. Arithmetic problems are offered in a story-telling way to identify mental alertness. It focuses upon attention and concentration while manipulating mental mathematical problems. For example, «Seven envelopes cost twenty five cents. How many envelopes can you buy if you have one dollar?».

– *Similarities in Dissimilar Objects.* A test taker is asked to find abstract similarities among different objects, for example among «a dog» and "a lion». With this test, abstract reasoning and power of conceptualization are measured.

Vocabulary. A participant is asked to explain the meaning of different words, for example "to crawl» or «a needle». The sub-test measures the comprehension of meanings and relations between the expressive words. For example, «What does the word zebra mean?»

As a result, the corpus contains the audio and textual material of 100 monologues, 56 dialogues and 100 verbal intelligence scores of the participants.

Language Style Matching. When a two-person conversation is kept going in a smooth and easy way, this means that the dialogue partners are trying to adapt to each other and to somehow coordinate their speech. The process of adaptation is based on synchronization with the emotional state of the other, listening to his or her point of view, finding proper words for expressing own thoughts and feeling and coordinating with his or her language style. Language style coordination may be reflected through the usage of similar words, phrases, sentences and sentence structures. In other words, if we analyze two texts with synchronized language styles and measure the similarity between them, its value should be high.

In this research, the degree of alignment between frequency distributions of a certain feature (token) was used as a measure of similarity between two texts. For compar-

ing the frequency distributions the chi-square test was chosen because it does not require the normality of distributions and is easy to implement. A detailed explanation of this method can be found in [5] and [6]. This technique has shown its efficiency in a number of studies, for example in analysis of authorship [7], political parties' activity [8, 9] and quantifying "strength of characterization within plays» [5]. In this section we will describe just the main idea of the approach.

Let F_i and F_j be two text files containing n_i and n_j tokens correspondingly. If F_i and F_j have the same language style, we consider the texts to be taken from the same population and the distributions of tokens from the two files should not be significantly different (null hypothesis).

The chi-square statistic is calculated based on the observed and expected values of tokens in both text-files. If the chi-value χ_i^2 is less than certain significance threshold c_i^2 (based on the degrees of freedom and a significance level), the null hypothesis is accepted and the two files may be considered as having a similar language style (making an assumption that the language style is reflected by tokens of this type). For estimating the degree to which the two texts were similar, we calculate the distance between these two values:

$$Similarity_i = S_i = \chi_i^2 - c_i^2 .$$

If $-c_i^2 \le S_i \le 0$, the similarity between the texts is significant. If $S_i > 0$, the null hypothesis is rejected: the analyzed texts have different language styles.

In this investigation four different types of tokens were used:

– *Letter n-gram distributions.*

– *Word n-gram distributions.*

– *Lemma n-gram distributions.* At first we analyzed all the lemmas which occurred in the monologues and the film. Further we will refer to this feature as *Lemma (Type 1)*. For taking into account that we work with spoken language, which may contain broken words, unfinished phrases and paraverbal expressions (like ah, hmm, etc.), for monologue analysis we used only lemmas which correspond to the following parts of speech: nouns, pronouns, verbs, adverbs, prepositions, conjunctions, interjections and articles. Such lemmas may be more important for reflecting language style matching. Let's refer to this feature as *Lemma (Type 2)*.

– *Part-of-speech n-gram distributions.* At first we analyzed n-gram distributions of all parts of speech occurred in the monologues *(Part-of-speech (Type 1))*. Secondly, n-grams were calculated only for parts of speech mentioned in the previous item *(Part-of-speech (Type 2))*.

Procedure and Results. In this research we analyzed differences in language styles of people with different verbal intelligence. The experiments described bellow will allow us to explore to what degree our test-persons matched the language style of the film when describing it and whether they were able to adapt to their dialogue partners or not.

Language Style Matching. Experiment 1. Using the k-means algorithm, the verbal intelligence scores of the test persons were partitioned into:

a) 2 clusters (Cluster P_1 consisted of test persons with lower verbal intelligence, P_2 contained candidates with higher verbal intelligence);

b) 3 clusters (P_1 - lower verbal intelligence, P_2 - average verbal intelligence, P_3 - higher verbal intelligence).

For each "couple» ($monol_i$ and $film$, $i = \overline{1,N}$, N - number of monologues) the similarity S_i among distributions of tokens (word, letter, lemma and part-of-speech n-grams, $n = \overline{1,10}$ was calculated. The mean values of S_i for each cluster were compared to each other using ANOVA. Features with significant ANOVA results for 2 clusters were:

– Word 3-g and 4-g distributions;

– Lemma (Type 1) 3-g and 4-g distributions;

– Lemma (Type 2) 3-g, 4-g, 5-g, 6-g and 7-g distributions;

– Part-of-speech (Type 1) 6-g and 7-g distributions;

– Part-of-speech (Type 2) 6-g, 7-g and 8-g distributions.

– Features with significant ANOVA results for 3 clusters were:

– Word 3-g distributions;

– Part-of-speech (Type 1) 6-g distributions;

– Part-of-speech (Type 2) 6-g distributions.

According to the results, letter n-gram distributions do not reflect differences in LSM. This feature may be more suitable for analyzing differences between two or more languages or determining whether two texts belong to the same author or not. In our experiments all the candidates spoke the same language and this feature occurred to be useless for estimating differences in language styles. For all the significant features, LSM of candidates with higher verbal intelligence was greater than of candidates with average and lower verbal intelligence. These results confirmed conclusions made in [3]: higher grade students tend to match the style of asked questions more than lower grade students.

Experiments with Dialogues. Experiment 1. For analyzing whether LSM depends on verbal intelligence of dialogue partners, all the dialogues were partitioned into the following groups (using verbal intelligence clusters obtained in the experiment with monologues (1a)):

a) L-L is a group of dialogues where both partners had lower verbal intelligence scores;

b) H-H is a group of dialogues where both partners had higher verbal intelligence scores;

c) L-H is a group of all the other dialogues.

For each dialogue, the similarity S_i was estimated: tokens' distributions of all the utterances of the first dia-

logue partner were compared with those of all the utterances of the second dialogue partner. Using ANOVA, the mean values of S_i of each group (L-L, H-H and L-H) were compared to each other. ANOVA did not show any significant differences.

Experiment 2. Analyzing the results from Experiment 1, we suggested that harmony in conversations may also depend on "closeness» of the dialogue partners. For example, two close friends may find hundreds of topics for their conversations. Our first question was whether they adapt to each other and synchronize their language during such discussions. Another question was whether any adaptation exists in conversations of people who see each other for the first time. For analyzing these situations, we used information about the level of acquaintance of the dialogue partners in our experiments and partitioned them into the following groups:

a) F-F is a group of dialogues with dialogue partners who were friends or relatives;

b) S-S is a group of dialogues with dialogue partners who had not met each other before the experiment (were strangers).

Again, the mean values of S_i for each group were compared to each other using ANOVA.

Features with significant ANOVA results were:

Word 3-g distributions;

– Lemma (Type 1) 3-g distributions;

– Lemma (Type 2) 3-g distributions;

– Part-of-speech (Type 1) 4-g and 5-g distributions;

– Part-of-speech (Type 2) 3-g, 4-g and 5-g distributions.

The distribution of these features showed that the similarities of language between friends or relatives were greater than between participants who had not met each other before.

Experiment 3. Our next purpose was to check whether verbal intelligence plays a certain role if we analyze dialogues between friends and strangers separately. For this purpose, ANOVA was applied to the mean values of S_i calculated for the following groups:

a) L-L, H-H and L-H only for dialogues between friends;

b) L-L, H-H and L-H only for dialogues between strangers.

In both cases ANOVA significant features were:

– Part-of-speech (Type 1) 6-grams;

– Part-of-speech (Type 2) 6-grams.

These features showed that dialogues between higher verbal intelligent participants had the highest similarity of language independent from whether the dialogue partners were friends or strangers. On the other hand, dialogues between lower verbal intelligent participants had the smallest value of LSM. Interestingly, LSMs of lower verbal intelligent friends were greater than LSMs of lower verbal intelligent strangers.

Analyzing the results we may say the degree to which a speaker adapts to his dialogue partner depends on the level of their acquaintance and the levels of their verbal intelligence. Of course other characteristics may influence on this process: openness to experience of the speakers, their mood, psychological states, etc. However, speakers' LSM may be used as a feature for improving automatic classification of speakers' verbal intelligence. On the other hand, the results suggest that the ability of a SLDS to match a user's language style and adapt to his verbal intelligence level should considerably improve its user-friendliness and attractiveness.

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

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

References

1. Ireland M. E., Slatcher R. B., Eastwick P. W., Scissors L. E., Finkel E. J., Pennebaker J. W. Language Style Matching Predicts Relationship Initiation and Stability. Psychological Science, XX(X), Sage Publications. 2010. P. 1–6,

2. Niederhoffer K. G., Pennebaker J. W. Linguistic Style Matching in Social Interaction // Journal of Language and Social Psychology. 2002. Vol. 21, № 4. P. 337–360.

3. Ireland M. E., Pennebaker J. W. Language Style Matching in Writing: Synchrony in Essays, Correspondence, and Poetry // Journal of Personality and Social Psychology. 2010. Vol. 99, № 3. P. 549–571.

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

5. Vogel C., Lynch G. Computational Stylometry: Who's in a Play? In Proceedings of COST 2102 Workshop (Patras)'2007. P. 169–186.

6. Straker D. Changing Minds. URL: http://changingminds.org/explanations/research/analysis/chisquare.htm.

7. Chaski C. Who Wrote It? Steps Toward a Science of Authorship Identification // Institute of Justice Journal. 233, P. 15-22.

8. Laver M. and Garry J. Estimating Policy Positions from Political Texts // American Journal of Political Science. 44(3)., P. 619-634.

9. Van Gijsel S., Vogel C. Inducing a Cline from Corpora of Political Manifestos / Aleksy M., et al. (eds.) // Proceedings of the International Symposium on Information and Communication Technologies. P. 304–310.

К. В. Заблоцкая, С. Г. Заблоцкий, Ф. Фернандес-Мартинес, В. Минкер

СОЧЕТАНИЕ ЛИНГВИСТИЧЕСКОГО СТИЛЯ И ВЕРБАЛЬНОГО ИНТЕЛЛЕКТА

Представлен сравнительный анализ лингвистических стилей людей, обладающих разным вербальным интеллектом. Для исследования был использован речевой корпус, состоящий из 100 монологов (пересказов одного и того же короткого фильма), 56 диалогов на одну тему и соответствующих оценок вербального интеллекта людей, участвующих в эксперименте. Согласно результатам, люди с более высоким вербальным интеллектом показали более близкое сходство лингвистического стиля при описании фильма и лучше адаптировались к партнерам по диалогу, чем испытуемые с меньшим вербальным интеллектом.

Ключевые слова: речевые диалоговые системы, лингвистический анализ, ANOVA.