Theme Identification in Telephone Service Conversations using Quaternions of Speech Features

Mohamed Morchid[1], Georges Linarès[1],
Marc El-Beze[1], Renato De Mori[1,2]

[1]LIA, University of Avignon, France
[2]McGill University, School of Computer Science, Montreal, Quebec, Canada
{mohamed.morchid,georges.linares,marc.elbeze}@univ-avignon.fr,
rdemori@cs.mcgill.ca

Abstract

The paper introduces new features for describing possible focus variation in a human/human conversation. The application considered is a real-life telephone customer care service. The purpose is to hypothesize the dominant theme of conversations between a casual customer calling. Conversations are processed by an automatic speech recognition system that provides hypotheses used for extracting word frequency. Features are extracted in different, broadly defined and partially overlapped, time segments. Combinations of each feature in different segments are represented in a quaternion algebra framework. The advantage of the proposed approach is made evident by the statistically significant improvements in theme classification accuracy.

Index Terms: Speech analytics, human/human conversation analysis, topic identification, quaternion algebra

1. Introduction

The application considered in this paper concerns the automatic analysis of telephone conversations [1] between an agent and a customer in the call center of the Paris public transport authority (RATP) [2]. The most important speech analytics for the application are the conversation themes. Relying on the ontology provided by the RATP, we have identified 8 themes related to the main reason of the customer call, such as *time schedules, traffic states, special offers, lost and found*,...

A conversation involves a customer, which is calling from an unconstrained environment (typically from train station or street, by using a mobile phone) and an agent which is supposed to follow a conversation protocol to address customer requests or complains. The conversation tends to vary according to the model of the agent protocol. This paper describes a theme identification method that relies on features related to this underlying structure of agent-customer conversation.

Here, the identification of conversation theme encounters two main problems. First, speech signals may contain very noisy segments that are decoded by an Automatic Speech Recognition (ASR) system. On such noisy environments, ASR systems frequently fail and the theme identification component has to deal with high Word Error Rates (WER $\simeq 58\%$).

Second, themes could be quite ambiguous, many speech acts being theme-independent (and sometimes confusing) due to the specificities of the applicative context: most of conversations evoke traffic details or issues, station names, time schedules, etc... Moreover, some of the dialogues contain secondary topics, augmenting the difficulty of dominant theme identification.

On the other hand, dialogues are redundant and driven by the RATP agents which try to follow, as much as possible, standard dialogue schemes. Considering that this underlying dialogue structure could bring relevant information about the conversation theme, we propose a conversation model based on the conversation split into four phases and on the quaternion algebra.

The quaternion algebra was introduced in [3], as an extension of the complex number field. A quaternion is composed of a real part and 3 imaginary parts. Their use has been proposed for modeling activities such as movements in computer vision, computer graphics and robotics. An interesting review with motivations and proposals for modeling interactive character motions can be found in [4]. In all these applications, quaternions provide a computationally effective formulation of object movement.

The proposed method consists in extending the classical vectorial model in which a document is represented by a vector of word frequencies [5]: here, a document is represented by a vector of quaternions, each quaternion grouping the four word frequencies estimated on each of the four parts of the conversation. The central intuition is that the quaternion model is better than other methods for combining the impact of different phases of the dialogue. In the next sections, we present in details the bases of this modeling scheme, its implementation in the experimental framework of the RATP call center and a step-by-step comparison to classical vectorial-model based system.

The rest of the paper is organized as follows. Next section briefly discusses the related work. Section 3 presents the proposed method. Section 4 describes the experiments and the results. Section 5 concludes and indicates the lines of future work.

2. Related works

Recent reviews for spoken conversation analysis, speech analytics, topic identification and segmentation can be found in [6, 7], [8], [9] and [10] respectively. Some important problems in finding topic dependent segments are the detection of segment boundaries and modeling the fact that segments may overlap. A generative model of lexical cohesion and using cue phrases for selecting samples in a guided search is proposed

This work was funded by the SUMACC and DECODA projects supported by the French National Research Agency (ANR) under contract ANR-10-CORD-007 and ANR-09-CORD-005.

in [11] for segmenting written text. A solution for considering statistical dependence among topic related segments is the semantic associative topic model (SATM) that is built on the notion of associated words in a probabilistic latent semantic analysis (pLSA) framework [12].

The model considered in this paper is not devoted to detect theme dependent segments, but to characterize a theme by the expression of its semantic components in possibly different phases of dialogue evolutions. Quaternions are proposed to represent features for this speech processing task. To our knowledge, the only application of quaternions to a speech processing task is for speech analysis with spatiotemporal Fourier descriptors [13].

3. Proposed approach

3.1. Theme-dependent bag-of-words

Here, we propose a quaternion-based representation that is supposed to capture the variations of word distributions in the conversation. This method will be compared to a classical one, based on bag-of-words (*BoW*) and vectorial representation of word frequencies.

BoW are basic features for topic identification. Their use, together with a comprehensive overview of features and methods for topic identification are described in [8]. All these competing approaches use the same set of theme-dependent *BoW* that are composed by estimating, on a training corpus, the discriminative capacity of words for each class. This discriminative score is the product of the word frequency (TF) [5], the inverse document frequency (IDF) [5] and the Gini purity criterion (GP, [14, 15]), defined as :

$$GP(w) = \sum_{i=1}^{N} P^2(t_i|w) \; .$$

Finally, theme-dependent *BoW* are obtained by selecting, for each of the N class t_i, the n-best words according to their discriminative capacity. The value n is estimated empirically on the development set and is the same for all classes.

3.2. Dialogue segmentation

The proposed method is motivated by the assumption that the structure of a dialogue offers an accurate point-of-view on the thematic structure, and helps to the theme recognition.

One of the key-point for capturing the structural features is the segmentation. Ideally, segmentation should match to the different dialogue phases but, in practice, these phases have a variable length, may overlap and the dialogues may not exhibit clear segment boundaries. Here, segmentation is used as a mean to extract features that will exhibit the focus variations along the document.

We tested two segmentation schemes (see figure 1). The first one consists in the intuitive left-right segmentation: the document is split into four successive overlapping segments. Each part covers 20% of the whole document, a segment is composed with two successive parts overlapping at 40%. The second one is a symmetric segmentation based on the position of words relatively to the centre of the dialogue. In this segmentation scheme, the four areas corresponds to (1) the first half of the document, (2) the middle area (50% centered), (3) the second half of the document and (4) the merge of the two extreme parts of dialogue (first 25% and last 25%). Figue 1 shows an example of document segmentation with the left-right (S_1, S_2, S_3, S_4) and the symmetric schemes (W_1, W_2, W_3, W_4).

These two segmentation strategies allow to extract four word statistics, that will be encapsulated into quaternions as described in the next section.

3.3. Quaternions representation

Following [3], a quaternion is an extension of a complex number uniquely defined in a four dimensional (4-D) space as a linear combination of four basis elements denoted as $1, i, j, k$. The element 1 is the identity element of the vector space. A quaternion Q is written as :

$$Q = r1 + xi + yj + zk \tag{1}$$

and represents a relation between the four real numbers r, x, y, z. In a quaternion, r is its real part while $xi + yj + zk$ is the imaginary part (I) or the vector part. There is a set of basic quaternion properties that are important for the further distance definition:

- all the possible products of i, j and k :

$$i^2 = j^2 = k^2 = ijk = -1 \tag{2}$$

- quaternion norm: $|Q| = \sqrt{r^2 + x^2 + y^2 + z^2}$
- normalized quaternion $Q^{\triangleleft}$

$$Q^{\triangleleft} = \frac{Q}{|Q|} \tag{3}$$

- inner product between two quaternions $Q = r1 + xi + yj + zk$ and $Q' = r'1 + x'i + y'j + z'k$

$$\langle Q, Q' \rangle = rr' + xx' + yy' + zz' \tag{4}$$

Given a segmentation $S = \{s_1, s_2, s_3, s_4\}$ of a document $d \in D$ and a vocabulary of words $v = \{w_1, \ldots, w_n, \ldots, w_{|v|}\}$, each word w_n is represented by the quaternion:

$$Q_d(w_n) = f_d^1(w_n)1 + f_d^2(w_n)i + f_d^3(w_n)j + f_d^4(w_n)k \tag{5}$$

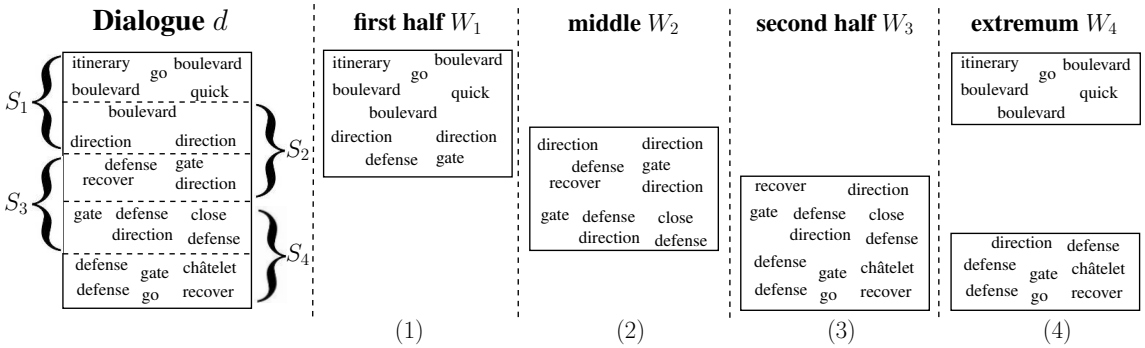
where $f_d^m(w_n)$ is the frequency of w_n in segment s_m of a document d.

More about hyper-complex numbers systems can be found in [16, 17, 18] and more precisely about quaternion in [19]. Figure 1 illustrates this quaternion-based representation of documents with the symmetric segmentation scheme. It shows a real dialogue from the test set, whose the theme was tagged as "itinerary".

A document $d \in D$ is represented by the N dimensional vector of quaternions $Q_d = [Q_d(w_n)]_{n \in N}$.

The equations (2) determine all the possible products of the bases. Given two quaternions represented by their real and imaginary parts as $Q_1 = r_1 + I_1$ and $Q_2 = r_2 + I_2$ basic operations are defined. If a quaternion is normalized (3) so that the sum of the squares of the four real numbers characterizing it sums to 1, then the quaternion represents an orientation and the distance between two quaternions of this type roughly corresponds to the angular distance of the two orientations represented by each quaternion. Let us consider two quaternions defined by the (5) and representing distributions of the same word w in two documents. Let $Q_{d_1}^{\triangleleft}$ and $Q_{d_2}^{\triangleleft}$ be the normalized versions (3) of the two quaternions Q_{d_1} and Q_{d_2}. The distance between the two documents follows from the double-angle ($cos(2\phi) = 2cos^2(\phi) - 1$) formula for cosine, together

Figure 1: Decoda test dialogue segmented according to left-right (S_1, S_2, S_3, S_4) and symmetric schemes (W_1, W_2, W_3, W_4). This dialogue was correctly labeled by the 2 quaternion-based approaches, since all TF-IDF-based approaches failed.



with the fact that the angle between orientations θ is precisely twice the angle ϕ ($\phi = \frac{\theta}{2}$) between two unit quaternions and:

$$cos(2\frac{\theta_{1,2}(w)}{2}) = 2cos^2(\frac{\theta_{1,2}(w)}{2}) - 1$$
$$= 2\left\langle Q^{\triangleleft}_{d_1}(w), Q^{\triangleleft}_{d_2}(w)\right\rangle^2 - 1$$

thus,

$$\theta_{1,2}(w) = cos^{-1}\left[2\left\langle Q^{\triangleleft}_{d_1}(w), Q^{\triangleleft}_{d_2}(w)\right\rangle^2 - 1\right]$$

where $\langle\cdot,\cdot\rangle$ is the inner product of two quaternions defined in (4). Computing the angular distance between the quaternions of each word for the two documents, a vector of distances is obtained. It represents the structural distortion between the two documents. Then, the similarity between two conversations is the mean score between each word of v :

$$\Theta(d_1, d_2) = \frac{1}{|v|}\sum_{w\in v}\theta_{1,2}(w) \quad (6)$$

3.4. Theme hypothesization

Our proposal, based on vectors of quaternions, is a representation paradigm rather than a classification method and various algorithms could be applied to make decision on quaternion feature space [20]. Therefore, we conducted contrastive experiments by using K-nearest-neighbors (KNN) [21] estimators, that require only a distance to be applied to non-standard features.

In our theme-identification task, KNN algorithm computes distortions between a test conversation C_i and each conversation in the train corpus. A subset $SS_k(C_i)$ of k nearest elements is extracted. The probability of the theme j given C_i is estimated by counting the number of conversations of this theme j in SS_k. In our experiments, the number of the nearest neighbors k is decided after empirical evaluations on the development set.

4. Experiments

4.1. Task and Corpus

Experiments were conducted on the Decoda corpus [2], composed by 1,242 telephone conversations from the call centre of the public transportation service in Paris. This corpus is split into a train set (740 conversations), a dev set (175 conversations) and a test set (327 conversations) and manually annotated with 8 conversation themes: *problems of itinerary, lost and found, time schedules, transportation cards, state of the traffic, fares, infractions* and *special offers*. Conversations have been manually transcribed and labeled with one theme label corresponding to the principal concern mentioned by the customer and are referred in the following with the label TRS.

4.2. ASR

The ASR system used for the experiments is derived from the LIA-Speeral Broadcast News system described in [22]. It relies on classical acoustic modeling with Hidden Markov Models, n-gram language models and an $A^\star$ search algorithm. To be effective on the Decoda setup, generic acoustic models were adapted with maximum a-posteriori probability (MAP) on 150 hours of speech in telephone bandwidth with the Decoda train set, and a specific 3-gram language model was estimated on the corresponding gold transcriptions, with a task-specific vocabulary of 5,782 words. An initial set of experiments was performed with this system, resulting to an overall WER on the test set of 58% (53% for agents and 63% for customers). These high error rates are mainly due to speech disfluencies and to adverse acoustic environments for some dialogues when, for example, users are calling from train stations or noisy streets with mobile phones. Furthermore, the signal of some sentences is saturated or of low intensity.

4.3. Baseline systems

We performed a step-by-step comparison of the quaternion-based representations, for the two proposed segmentation schemes, with six contrastive systems. All are compared by using strictly the same experimental setup, including a common bag-of-word set, common classification method and, naturally, the Decoda train, test, and dev corpus.

The two quaternion-based systems evaluated use left-right and symmetric segmentation schemes, as described in the last section. They are denoted *LRQ* and *SSQ* respectively.

The first baseline system is based on the classical TF-IDF based approach, in which each document is characterized by a vector of word frequencies. Frequencies are estimated on the whole conversation, and the cosine metric is used for document-to-document distance.

An important aspect of our proposal is the segmentation

Table 1: **Results of theme classification accuracy for the baseline systems** (Confidence interval of $\pm 4.43\%$ for the M4D system)

DATA		*ACCURACY (%)*									
test	**train**	TS_1		TS_2		TS_3		TS_4		**M4D**	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Trs	Trs	*70.3*	64.4	*69.1*	62.5	*68*	62.2	*61.1*	39.8	*86.2*	78.8
Asr	Trs	*65.1*	55	*66.8*	54.1	*61.7*	48.3	*53.7*	39.1	*81.7*	70
Asr	Asr	*64*	56.2	*65.7*	50.7	*63.4*	48.3	*53.1*	44.3	*78.1*	67.4
Asr	A.+T.	*64.5*	57.1	*66.2*	50.7	*61.7*	50.1	*55.4*	45.5	*82.5*	70.3

step, that is based on the intuition that position-dependent distributions bring something characteristic of the theme. The proposed system combines features dependent from the dialogue phases (or segments), and a modeling paradigm based on the quaternion algebra.

In order to evaluate separately the potential gains obtained by these two aspects of our proposal, we tested a TF-IDF system operating on frequency-vectors estimated on the four successive segments of conversation rather than on the whole dialogue. In this system (named *M4D* in the next), a vector representing a document includes $4 * n$ coefficients, for a n-word *BoW*, cosine distance being used for the classification step. This method is directly comparable to the quaternion-based approach that uses the same low-level feature set (SSQ), which is composed by segment-dependent word frequencies. It aims at evaluating the specific interest of quaternions, independently from the basic feature set resulting from segmentation.

Since the later approach groups all segment-dependent frequencies in a large vector, we performed a last test where we estimated the individual contribution of each of the four conversation parts of the left-right segmentation scheme. This is achieved by evaluating the performance of TF-IDF systems using only word distributions computed on only one of the conversation parts. These systems are named TS_1, TS_2, TS_3 and TS_4.

All these systems are tested with KNN classifiers operating with train set composed by manual transcription of train dialogues (TRS), automatically transcription of dialogues (ASR), and the both (ASR+T). For comparison, we present results on gold trancriptions of the test set and the realistic case of identification from ASR outputs (ASR).

Table 2: **Results of theme classification accuracy** (Confidence interval of $\pm 4.56\%$ for the SSQ system)

DATA		*ACCURACY (%)*							
test	**train**	**TF-IDF**		**M4D**		**LRQ** Θ		**SSQ** Θ	
		Dev	Test	Dev	Test	Dev	Test	Dev	Test
Trs	Trs	72	71.7	86.2	78.8	89	85.3	92.2	87.4
Asr	Trs	68.4	64.4	81.7	70	82.2	74.1	84.6	76
Asr	Asr	65	61.3	78.1	67.4	80.5	71.7	82.2	73.9
Asr	A.+T.	67.5	62.2	82.5	70.3	81	73.3	83.4	**76.9**

4.4. Results and discussion

Previous experiments were conducted in the Decoda framework, mainly on smaller data sets. *BoW* were used for the same purpose in previous experiments on data belonging to a portion of the actual corpus [23, 24]. In [24], phrases were added to word features and used in a cosine similarity measure. In [23], *BoW* are proposed as features for a multiple view *AdaBoost* based topic classification approach. Five separate views are introduced for respectively the agent and the customer, the dialogue turn boundaries, the durations and the name entities. Our experiments rely on an augmented test set, containing complex and diversified samples.

Table 2 reports results obtained by TF-IDF, M4D and the two quaternion-based systems using left-rigth (LRQ) and symmetric (SSQ) segmentation schemes. We first observe that dialogue segmenting yields to a significant improvement of accuracy: M4D system outperforms TF-IDF by about 7% absolute, validating the initial intuition of the relevance of an underlying conversation structure for theme identification.

Quaternions provide additional gain of 6.5% for LRQ, and 8.6% for SSQ segments. This result confirms that, at contrary to classical matrix representations as included in M4D system, quaternions allow to capture dependencies of word distributions belonging to the document. These dependencies seem clearly relevant for theme extraction.

The difference between LRQ and SSQ segmentations are limited but relatively unexpected; this result may be due to the fact that the latent structure of conversation could not match, as well as expected, the proposed 4-part scheme. The comparison of single-segment systems (see table 1) exhibits a slight advantage of the first segment, but individual performances are far from the one obtained with combined features (M4D, LRQ and SSQ). Globally, these measures suggest that other segmentation approaches could still improve overall system performance.

Finally, the use of both automatic and gold transcriptions (A.+T. in the table 2) yields to significant improvements, in comparison to KNN applied to only one of the two transcription sources (ASR or TRS). The merging of the two set of examples seems to improve the system robustness to ASR errors.

5. Conclusion and future work

A new method based on quaternions has been introduced for integrating features related to different segments of a conversation with the purpose of hypothesizing problem types expressed by a customer to an agent. It consisted in segmenting the conversation, in extracting low-level features from each parts and in combining them into hypercomplex numbers.

Results have validated both the idea of capturing focus variation along the conversation, and the effectiveness of the quaternion based coding of small set of word-dependent features. In comparison to high dimensionality matrix representations, quaternions allow to model feature dependencies, using metrics based on minimal distortion. Our results have shown the significant gain provided by the quaternion-based features on the targeted task, on which feature dependencies were expected to be strong.

Research will continue mostly along three lines, namely identification of possible multiple themes, definition of suitable confidence indicators based on quaternion features and the search of relevant segmentation schemes.

Finally, the proposed representation paradigm could be applied to different types of structured documents, in various contexts. We plan now to evaluate the genericity of this method by applying it to various natural language processing tasks.

6. References

[1] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *in IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 1992*, vol. 1. IEEE, pp. 517–520.

[2] F. Bechet, B. Maza, N. Bigouroux, T. Bazillon, M. El-Beze, R. De Mori, and E. Arbillot, "Decoda: a call-centre human-human spoken conversation corpus," in *inInternational Conference on Language Resources and Evaluation*. LREC, 2012.

[3] S. Hamilton, *Elements of quaternions*. Longmans, Green, & co., 1866.

[4] M. Johnson, "Exploiting quaternions to support expressive interactive character motion," Ph.D. dissertation, Massachusetts Institute of Technology, 2002.

[5] G. Salton, "Automatic text processing: the transformation," *Analysis and Retrieval of Information by Computer*, 1989.

[6] K. Lagus and J. Kuusisto, "Topic identification in natural language dialogues using neural networks," in *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 95–102. [Online]. Available: http://www.aclweb.org/anthology/W02-1014

[7] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*. Wiley, 2011.

[8] T. Hazen, "Topic identification," *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pp. 319–356, 2011.

[9] I. Melamed and M. Gilbert, "Speech analytics," *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pp. 397–416, 2011.

[10] M. Purver, "Topic segmentation," *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pp. 291–317, 2011.

[11] J. Eisenstein and R. Barzilay, "Bayesian unsupervised topic segmentation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2008, pp. 334–343.

[12] M. Wu, H. Lee, and H. Wang, "Exploiting semantic associative information in topic modeling," in *in ISCA/IEEE Workshop on Spoken Language Technology (SLT), 2010*. IEEE, 2010, pp. 384–388.

[13] E. Bayro-Corrochano, N. Trujillo, and M. Naranjo, "Quaternion fourier descriptors for the preprocessing and recognition of spoken words using images of spatiotemporal representations," *Journal of Mathematical Imaging and Vision*, vol. 28, no. 2, pp. 179–190, 2007.

[14] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. Chapman & Hall/CRC, 1984.

[15] T. Dong, W. Shang, and H. Zhu, "An improved algorithm of bayesian text categorization," *Journal of Software*, vol. 6, no. 9, pp. 1837–1843, 2011.

[16] I. Kantor, A. Solodovnikov, and A. Shenitzer, *Hypercomplex numbers: an elementary introduction to algebras*. Springer-Verlag, 1989.

[17] J. B. Kuipers, *Quaternions and rotation sequences*. Princeton university press Princeton, NJ, USA:, 1999.

[18] F. Zhang, "Quaternions and matrices of quaternions," *Linear algebra and its applications*, vol. 251, pp. 21–57, 1997.

[19] J. Ward, *Quaternions and Cayley numbers: Algebra and applications*. Springer, 1997, vol. 403.

[20] B. Harish, D. Guru, and S. Manjunath, "Representation and classification of text documents: A brief review," *International Journal of Computer Applications IJCA*, no. 2, pp. 110–119, 2010.

[21] J. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[22] G. Linarès, P. Nocéra, D. Massonie, and D. Matrouf, "The lia speech recognition system: from 10xrt to 1xrt," in *Text, Speech and Dialogue*. Springer, 2007, pp. 302–308.

[23] S. Koço, C. Capponi, and F. Béchet, "Applying multiview learning algorithms to human-human conversation classification," in *in International conference of the Speech Communication Association (InterSpeech) 2012*, 2012.

[24] B. Maza, M. El-Beze, G. Linares, and R. Mori, "On the use of linguistic features in an automatic system for speech analytics of telephone conversations," in *in International conference of the Speech Communication Association (InterSpeech) 2011*, 2011.