Multiple topic identification in telephone conversations

Xavier Bost[1], Marc El-Beze[1], and Renato De Mori[1,2]

[1]LIA, University of Avignon, France
[2]McGill University, School of Computer Science, Montreal, Quebec, Canada
{xavier.bost, marc.elbeze, renato.demori}@univ-avignon.fr

Abstract

This paper deals with the automatic analysis of conversations between a customer and an agent in a call centre of a customer care service. The purpose of the analysis is to hypothesize themes about problems and complaints discussed in the conversation. Themes are defined by the application documentation topics. A conversation may contain mentions that are irrelevant for the application purpose and multiple themes whose mentions may be interleaved portions of a conversation that cannot be well defined. Two methods are proposed for multiple theme hypothesization. One of them is based on a cosine similarity measure using a bag of features extracted from the entire conversation. The other method introduces the concept of thematic density distributed around specific word positions in a conversation. In addition to automatically selected words, word bigrams with possible gaps between successive words are also considered and selected. Experimental results show that the results obtained with the proposed methods outperform the results obtained with support vector machines on the same data. Furthermore, using the theme skeleton of a conversation from which thematic densities are derived, it will be possible to extract components of an automatic conversation report to be used for improving the service performance.

Index Terms: multi-topic audio document classification, human/human conversation analysis, speech analytics, distance bigrams

1. Introduction

There has been a growing interest in recent years on speech technology capabilities for monitoring telephone services. In order to provide a high level of efficiency and user satisfaction, there is a consensus on the importance of improving systems for analysing human/human conversations to obtain reports on customer problems and the way an agent has solved them. With these reports, useful statistics can be obtained on problem types and facts, the efficiency of problem solutions, user attitude and satisfaction.

The application considered in this paper deals with the automatic analysis of dialogues between a call centre agent who can solve problems defined by the application domain documentation and a customer whose behaviour is unpredictable. The customer is expected to seek information and/or formulate complaints about the Paris transportation system and services.

The application documentation is a concise description of problem themes and basic facts of each theme considered as a conversation topic. The most important speech analytics for the application are the themes even if facts and related complementary information are also useful.

A conversation may contain more than one semantically related themes. Depending on the type of relation, some themes

discussed in a conversation may be irrelevant for the application task. For example, a customer may inquiry about an object lost on a transportation mean that was late. In such a case, the loss is a much more relevant theme than the traffic state. A component of the automatic analysis is an Automatic Speech Recognition (ASR) system that makes recognition errors. In the above example, failure to detect the mention to traffic state can be tolerated if the mention to the loss fact is correctly hypothesized but in general, all the mentioned themes must be taken into account. This paper focuses on the detection of relevant themes in a service dialogue. Conversations to be analysed may be about one or more domain themes. Detecting the possibility of having multiple themes and identifying them is important for estimating proportions of customer problems involving related topics. Different themes can be mentioned in disjoint discourse segments. In some cases, mentions of different themes may coexist in short segments or even in a single sentence. Even when different themes are mentioned in non-overlapping discourse segments, segment boundaries may be difficult to estimate because of the errors introduced by the ASR system and the imprecise knowledge about the language structures used by casual users in the considered real-world situation. In spite of the just discussed difficulties, it is worth considering the possibility of extracting suitable features for theme detection considering that it is likely to have several mentions of a theme in a conversation.

The paper is structured as follows. Section 2 of the paper discusses related work. Section 3 introduces the application domain and the features used for theme hypothesization. The first of two approaches to multiple theme identification is introduced in section 4. It is based on the automatic estimation of decision parameters applied to global cosine similarity measures. The second approach, described in section 5, introduces the concept of dialogue skeleton based on which thematic densities of features can be computed and used as soft detections of possibly overlapping locations in a dialogue where one or more themes are mentioned. Decision making with this approach is also described. Experimental results are presented in Section 6.

2. Related work

Human/human spoken conversation analyses have been recently reviewed in [1]. Methods for topic identification in audio documents are reviewed in [2]. Solutions for the detection of conversation segments expressing different topics have been proposed in many publications, recently reviewed in [3]. Interesting solutions have been proposed for linear models of non-hierarchical segmentation. Some approaches propose inference methods for selecting segmentation points at the local maxima of cohesion functions. Some functions use features extracted in each conversation sentence or in a window including a few sen-

tences. Some search methods detect cohesion in a conversation using language models and some others consider hidden topics shared across documents.

A critical review on lexical cohesion can be found in [4] who propose an unsupervised approach for hypothesizing segmentation points using cue phrases automatically extracted from unlabelled data. An evaluation of coarse-grain segmentation can be found in [5].

Multi-label classification is discussed in [6] and [7] mostly for large collections of text documents. Particularly interesting is a technique called creation consisting in creating new composite labels for each association of multiple labels assigned to an instance.

This paper extends concepts found in the recent literature by introducing a version of bigram words that may have a gap of one word. These features are used in decision strategies based on a constrained application of the cosine similarity and on a new definition of soft theme density location in a conversation.

3. Application task and features

The application task is about a customer care service (CCS). The purpose of the application is to monitor the effectiveness of a call centre by evaluating proportions of problem items and solutions. Application relevant information is described in the application requirements focussing on an informal description of useful speech analytics not completely defined. The requirements and other application documentation essentially describe the types of problems a CCS agent is entitled to solve.

This paper proposes a new approach for automatically annotating dialogues between an agent and a customer with one or more application theme labels belonging to the set $\mathbb{C}$ defined as follows:

$\mathbb{C} := \{$ itinerary, lost and found, time schedules, transportation card, traffic state, fine, special offers $\}$.

Given a pair (d, c), where $d \in \mathbb{X}$, a spoken dialogue in the corpus $\mathbb{X}$, is described by a vector v_d of features and $c \in \mathbb{C}$ is a class corresponding to a theme described by a vector v_c of features of the same type as v_d, two classification methods, indicated as γ_1 and γ_2, are proposed for multiple theme classification.

With the purpose of increasing the performance of automatic multiple theme hypothesization, bigrams were added to the lexicon of 7217 words with the possibility of having also distance bigrams made of pairs of words distant a maximum of two words. The feature set increases to 160433 features with this addition. In order to avoid data sparseness effects, a reduced feature set V was obtained by selecting features based on their purity and coverage. Purity of a feature t is defined with the Gini criterion as follows:

$$G(t) = \sum_{c \in \mathbb{C}} \mathbb{P}^2(c|t) = \sum_{c \in \mathbb{C}} \left(\frac{df_c(t)}{df_{\mathbb{T}}(t)} \right)^2 \qquad (1)$$

where $df_{\mathbb{T}}(t)$ is the number of dialogues of the train set $\mathbb{T}$ containing term t and $df_c(t)$ is the number of dialogues of the train set containing term t in dialogues annotated with theme c.

A score $w_c(t)$ is introduced for feature t in the entire train set collection of dialogues discussing theme c. It is computed as follows:

$$w_c(t) = df_c(t).idf^2(t).G^2(t) \qquad (2)$$

where $idf(t)$ is the inverse document frequency for feature t.

4. Using a global cosine similarity measure

The classical cosine measure of similarity between the two vectors v_d $(d \in \mathbb{X})$ and v_c $(c \in \mathbb{C})$ is defined as:

$$\mathrm{sc}(d, c) = \cos(\widehat{v_d, v_c}) = \frac{\sum_{t \in d \cap c} w_d(t).w_c(t)}{\sqrt{\sum_t w_d(t)^2.w_c(t)^2}} \qquad (3)$$

where $w_d(t)$ is a score for feature t in dialogue d.

Let $\gamma_1(d)$ be the set of themes discussed in dialogue d. A first decision rule for automatically annotating dialogue d with a theme class label c is:

$$c \in \gamma_1(d) \implies \mathrm{sc}(d, c) \geqslant \rho.\mathrm{sc}(d, \hat{c}) \qquad (4)$$

where $\hat{c} := \arg\max_{c' \in \mathbb{C}} \mathrm{sc}(d, c')$; and $\rho \in [0; 1]$ is an empirical parameter whose value is estimated by experiments on the development set.

If the score of $\hat{c}$ is too low, then the application of the above rule is not reliable. To overcome this problem, the following additional rule is introduced:

$$c \in \gamma_1(d) \implies \mathrm{sc}(d, c) \geqslant v \sum_{c' \in \mathbb{C}} \mathrm{sc}(d, c') \qquad (5)$$

where $v \in [0; 1]$ is another parameter whose value is estimated by experiments on the development set.

4.1. Parameter estimation

The values of parameters ρ and v have been estimated using 20 subsets of 98 dialogues each belonging to the development set and selected with the same proportion of single and multiple theme dialogues as in the development set. In order to estimate the optimal values $\hat{\rho}$ and $\hat{v}$ of these two parameters, the following decision rule has been applied:

$$(\hat{\rho}, \hat{v}) = \arg\max_{(\rho, v) \in [0;1]^2} \sum_{i=1}^{20} F(\gamma_{1(\rho, v)}, \mathbb{D}_i) \qquad (6)$$

where $F(\gamma_{1(\rho, v)}, \mathbb{D}_i)$ is the F-score (defined in the following subsection) computed using model γ_1 with the estimated values of ρ and v in the subset $\mathbb{D}_i \subset \mathbb{D}$.

The optimal value of ρ, the proportion of the highest score required for assigning themes to a dialogue, has been evaluated to $\hat{\rho} = 0.69$, and the optimal value of v, the threshold required, to $\hat{v} = 0.16$.

4.2. Performance measures

The proposed approaches have been evaluated following procedures discussed in [7] with measures used in Information Retrieval (IR) and accuracy as defined in the following for a corpus $\mathbb{X}$ and a decision strategy γ.

Of particular importance is the F-score: based on this measure, and as mentioned in subsection 6.3, it is possible to find the best trade-off between precision and recall by rejecting some of the dialogues.

Recall :

$$R(\gamma, \mathbb{X}) = \frac{1}{|\mathbb{X}|} \sum_{d \in \mathbb{X}} \frac{|\gamma(d) \cap L(d)|}{|L(d)|} \qquad (7)$$

where $L(d)$ indicates the set of theme labels annotated for conversation d.

Precision :

$$P(\gamma, \mathbb{X}) = \frac{1}{|\mathbb{X}|} \sum_{d \in \mathbb{X}} \frac{|\gamma(d) \cap L(d)|}{|\gamma(d)|} \qquad (8)$$

F-score :

$$F(\gamma, \mathbb{X}) = \frac{2P(\gamma, \mathbb{X})R(\gamma, \mathbb{X})}{P(\gamma, \mathbb{X}) + R(\gamma, \mathbb{X})} \qquad (9)$$

Accuracy :

$$A(\gamma, \mathbb{X}) = \frac{1}{|\mathbb{X}|} \sum_{d \in \mathbb{X}} \frac{|\gamma(d) \cap L(d)|}{|\gamma(d) \cup L(d)|} \qquad (10)$$

5. Automatic annotation based on thematic densities

5.1. Thematic density

The contribution to theme c of the features at the i-th location in a dialogue is:

$$w_c(p_i) = \frac{1}{\| v_c \|} \sum_{t \in T_{p_i}} w_c(t) \ \ (i = 1, ..., n) \qquad (11)$$

where T_{p_i} is the set made of the i-th word in a conversation and the bigrams associated with it.

A thematic density $d_c(p_i)$ of theme c is associated with position i and is defined as follows:

$$d_c(p_i) = \frac{\sum_{j=1}^{n} \frac{w_c(p_j)}{\lambda^{d_j}}}{\sum_{j=1}^{n} \frac{1}{\lambda^{d_j}}} \ \ (i = 1, ..., n) \qquad (12)$$

where $\lambda \geqslant 1$ is a parameter of sensitivity to proximity whose value is estimated by experiments on the development set and $d_j := |i - j|$.

5.2. Dialogue skeleton

The theme density at a specific dialogue location makes it possible to derive a thematic skeleton of a dialogue.

Figure 1 shows the skeleton of a dialogue obtained with an automatic transcription for $\lambda = 1$ and $\lambda = 1.05$ (fig 1- a) and $\lambda = 2.8$ (fig 1- b). The figures plot the thematic density as function of the dialogue location measured in numbers of words preceding the location. The conversation is about a request of bus *schedule* (indicated as HORR) and a *fare* (indicated as TARF). Three functions are plotted. Two for these themes and a third for the theme *itinerary* (indicated as ITNR).

As an example, an excerpt from the dialogue obtained from manual transcriptions is reported in the following. The dialogue positions corresponding to each turn are in brackets.

– *Customer* [82–100] : I would like to know if busses start running in early morning.
– *Agent* [101–118] : First start is at 5 :45 at Opera square.
– (...)
– *Agent* [304–314] : It will be there at 8 :48.
– (...)
– *Agent* [442–470] : eh no .... There are specific fares for the airport. I'll give you the amount .... Fare is 9 euros 10.
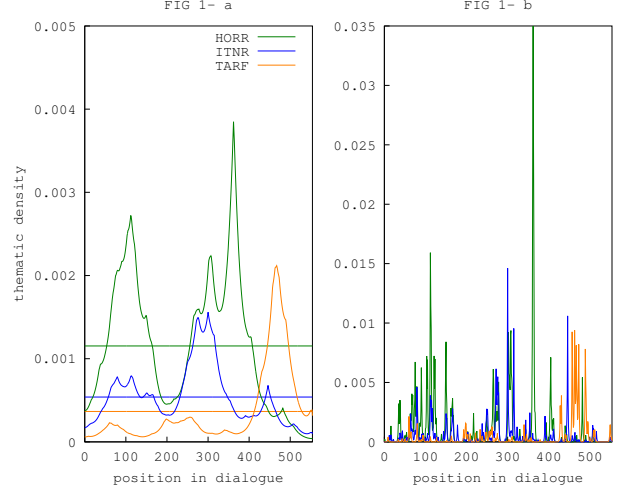– *Customer* [471–496] : 9 euros 10, should we pay 9 euros 10 cash.

Figure 1: *Thematic densities as function of location in a dialogue skeleton. Densities are plotted for $\lambda = 1$ and $\lambda = 1.05$ (fig 1- a) and $\lambda = 2.8$ (fig 1- b) and three themes: schedule (indicated as* HORR*), fare (indicated as* TARF*) and itinerary (indicated as* ITNR*.*

With $\lambda = 2.8$ (fig 1- b), distant contexts tend to be neglected. In this case decision may be adversely affected by isolated features that may not be relevant for theme hypothesization. For example, the expression *there* ("la-bas" in French) in turn [304–314] tends to show more evidence for itinerary with a peak of density at location 300, while it is used here in a request of schedule.

Local context is more appropriately taken into account with $\lambda = 1.05$ (fig 1- a) with the result of reducing the relevance of the itinerary hypothesis just supported by the word *there*. When $\lambda = 1$ (horizontal lines in fig 1- a), close contexts tend to be neglected giving more importance to global features used in the approach introduced in Section 4.

In conclusion, with a well-suited value of λ (=1.05), thematic coherence tends to make decisions more accurate.

5.3. Decision making

A theme c is considered as discussed in a dialogue d if it has dominant density at a location of the dialogue (rule (13)) and if the sum of its densities in positions where it is dominant exceeds an empirically determined threshold (rule (14)) :

$$c \in \gamma_2(d) \Rightarrow \exists i \in [\![1; n]\!] : \forall c' \in \mathbb{C}, \ d_c(p_i) \geqslant d_{c'}(p_i) \quad (13)$$

$$c \in \gamma_2(d) \Longrightarrow \sum_{i \in I} d_c(p_i) \geqslant v \sum_{i=1}^{n} d_{c_i}(p_i) \qquad (14)$$

where $v \in [0, 1]$ is a parameter whose value is estimated by experiments on the developement set; c_i is the theme of dominant density at the i-th position in the dialogue; and $I := \{n \in \mathbb{N} \mid \forall c' \in \mathbb{C}, \ d_c(p_n) \geqslant d_{c'}(p_n)\}$.

6. Experiments

6.1. Experimental framework

Experiments have been performed using an ASR system described in [8]. It is based on triphone acoustic hidden Markov models (HMM) with mixtures of Gaussians belonging to a set of 230000 distributions. Model parameters were estimated with maximum *a posteriori* probability (MAP) adaptation of 150 hours of speech in telephone bandwidth with the data of the train set. A corpus of 1658 telephone conversations was collected at the call centre of the public transportation service in Paris. The corpus is split into a train, a development and a test set containing respectively 884, 196 and 578 conversations. A 3-gram language model (LM) was obtained by adapting with the transcriptions of the train set a basic LM. An initial set of experiments were performed with this system resulting with an overall WER on the test set of 57% (52% for agents and 62% for users). 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 due to the distance between speakers and phones.

The annotation with possible multiple themes of the development and test corpora has been performed in a batch process by maximizing the agreement between three annotators.

It is important to notice that the training corpus dialogues have been labelled on the fly by the agents with the constraint to choose one and only one theme corresponding to the main customer concern. When this was not clear, the annotation was based on the problem expressed at the beginning of the conversation. With such a procedure it is not possible to create bi-labels as described in [6] and [7] and mentioned in Section 2.

6.2. Evaluation

For the sake of comparison, the results obtained with the proposed classification approaches have been compared with the results obtained with a support vector machine (SVM) using the same features (unigrams and bigrams with possible gap of one word) and a linear kernel. For every theme $c_i \in \mathbb{C}$, a binary classifier $\gamma_i : \mathbb{X} \to \{c_i, \overline{c_i}\}$ is defined. Every pair $(d, c_i) \in \mathbb{X} \times \mathbb{C}$ is associated with the score computed by this classifier. The candidate theme hypotheses for a conversation are those whose score is in an interval corresponding to an empirically determined proportion of the highest one. In addition to that, the hypothesis with the highest score must be above a threshold empirically determined for this purpose.

Results obtained with the cosine measure and with the theme density are reported in Table 1 for the development set and in Table 2 for the test set. MAN and ASR respectively indicate manual transcriptions and automatic transcriptions obtained with the most likely sequence of word hypotheses generated by the ASR system.

6.3. Results analysis

The development set was collected in the same time period (fall) as the train set, while the test set was collected in the summer. The difference in the results obtained with the test and the development sets can be explained in part considering the frequency of different events in the two time periods (*e. g.* strikes in the fall and specific maintenance works in the summer).

The strategies of all the used methods for hypothesizing a theme in addition to the dominant one give better results on the

Table 1: *Results obtained with the* SVM*, the cosine measure and the theme density for the development set.*

DEV	MAN			ASR		
	SVM	cos.	dens.	SVM	cos.	dens.
Accuracy	0.77	0.85	0.85	0.70	0.80	0.81
Precision	0.88	0.92	0.94	0.79	0.87	0.90
Recall	0.85	0.92	0.88	0.81	0.89	0.86
F-score	0.86	**0.92**	0.91	0.80	**0.88**	**0.88**

Table 2: *Results obtained with the* SVM*, the cosine measure and the theme density for the test set. For the cosine-based method, the estimated confidence interval is* ± *2.74% for manual transcriptions, and* ± *3% for the* ASR *output.*

TEST	MAN			ASR		
	SVM	cos.	dens.	SVM	cos.	dens.
Accuracy	0.74	0.78	0.78	0.64	0.71	0.71
Precision	0.85	0.86	0.86	0.72	0.79	0.79
Recall	0.83	0.87	0.85	0.77	0.83	0.80
F-score	0.84	**0.87**	0.85	0.75	**0.81**	0.80

hypothesized latter theme compared to those obtained for the former one (monolabel categorization). For example, using the relative value of the cosine measure with the dev set, improvements from 0.88 to 0.92 using the manual transcriptions and from 0.85 to 0.88 with the automatic transcriptions are respectively observed for the F-score. The same improvements are observed with the test set.

Using the development set for inferring a rejection rule based on close scores between the first two hypotheses, an F-score of 0.83 is obtained on the automatic transcriptions of the test set with a rejection rate close to 10%, the rate of disagreement between human annotators.

7. Conclusion and future work

Features have been proposed for the hypothesization of one or more themes mentioned in a conversation between a call centre agent and a calling customer. They are sets of words, bigrams and distant bigrams automatically selected in the application domain.

Two approaches have been proposed for multiple theme hypothesization. A first approach is based on a cosine similarity measure applied to the features extracted from a conversation and the other based on a new definition of theme density obtained considering a conversation skeleton. The approaches have been evaluated and have shown to outperform an SVM classifier using the same features and data.

Future research will include the search for confidence measures suitable for the multi topic task and the use of conversation skeletons for extracting short reports on agent/customer dialogues. From these reports, proportions of speech analytics will be extracted and used for monitoring frequencies, importance and solution rates for different type of facts and problems.

8. Acknowledgements

9. References

[1] Tur, G. and Hakkani-Tur D., "Human/human conversation understanding", ch. 9 of (G. Tur and R. De Mori), pp. 228–255, J. Wiley, 2011.

[2] Hazen, T. J., "Topic identification", ch. 12 of (G. Tur and R. De Mori), pp. 319–356, J. Wiley, 2011.

[3] Purver M., "Topic segmentation", ch. 11 of (G. Tur and R. De Mori), pp. 290–317, J. Wiley, 2011.

[4] Eisenstein, J. and Barzilay, R., "Bayesian Unsupervised Topic Segmentation", EMNLP, 2008, pp. 334–343.

[5] Niekrasz, J. and Moore, J. D., "Unbiased discourse segmentation evaluation", 2010 IEEE Spoken Language Technology workshop, Berkeley, CA, Dec 2010, pp. 43–48.

[6] Carvalho, A. and Freitas, A., "A tutorial on multi-label classification techniques", Foundations of Computational Intelligence, Vol. 5 of Studies in Computational Intelligence 205, pp. 177–195, Springer, September 2009.

[7] Tsoumakas, G. and Katakis, I., "Multi-label classification: an overview", International Journal of Data Warehousing and Mining, 3(3): pp. 1–13, 2007.

[8] Linares, G., Nocera, P., Massonie, D. and Matrouf, D., "The LIA speech recognition system: from 10xRT to 1xRC", International Conference on Speech, Text and Dialogue, Pilsen, Tcheck Republic, 2007, Lecture Notes in Computer Science, volume 4629/2007, pp. 302–308.

[9] Maza, B., El-Beze, M., Linares, G. and De Mori, R., "On the use of linguistic features in a automatic system for speech analytics of telephone conversations", Interspeech 2011, pp. 2049–2052.

[10] Koco, S., Capponi, C. and Bechet, F., "Applying multiview learning algorithms to human-human conversation classification", Interspeech 2012, Portland.

[11] Tur, G. and De Mori, R. Eds, "Spoken Language Understanding: Systems for Extracting Semantic Information from Speech", J. Wiley, 2011.