



**HAL**  
open science

# Concept Discovery and Automatic Semantic Annotation for Language Understanding in an Information-Query Dialogue System Using Latent Dirichlet Allocation and Segmental Methods

Nathalie Camelin, Boris Detienne, Stéphane Huet, Dominique Quadri, Fabrice  
Lefèvre

► **To cite this version:**

Nathalie Camelin, Boris Detienne, Stéphane Huet, Dominique Quadri, Fabrice Lefèvre. Concept Discovery and Automatic Semantic Annotation for Language Understanding in an Information-Query Dialogue System Using Latent Dirichlet Allocation and Segmental Methods. Fred A., Dietz J.L.G., Liu K., Filipe J. Knowledge Discovery, Knowledge Engineering and Knowledge Management, 348, Springer, pp.45-59, 2013, Communications in Computer and Information Science, 978-3-642-37185-1. 10.1007/978-3-642-37186-8\_3 . hal-01314550

**HAL Id: hal-01314550**

**<https://hal.science/hal-01314550v1>**

Submitted on 27 Feb 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Concept discovery and automatic semantic annotation for language understanding in an information-query dialogue system using Latent Dirichlet Allocation and segmental methods

Nathalie Camelin<sup>1</sup>, Boris Detienne<sup>2</sup>, Stéphane Huet<sup>2</sup>, Dominique Quadri<sup>2</sup>, and Fabrice Lefèvre<sup>2</sup>

<sup>1</sup> Université du Maine, LIUM, Le Mans, France  
forename.surname@lium.univ-lemans.fr,

<sup>2</sup> Université d'Avignon, LIA-CERI, Avignon, France  
forename.surname@univ-avignon.fr

**Abstract.** Efficient statistical approaches have been recently proposed for natural language understanding in the context of dialogue systems. However, these approaches are trained on data semantically annotated at the segmental level, which increases the production cost of these resources. This kind of semantic annotation implies both to determine the concepts in a sentence and to link them to their corresponding word segments. In this paper, we propose a two-step automatic method for semantic annotation. The first step is an implementation of the latent Dirichlet allocation aiming at discovering concepts in a dialogue corpus. Then this knowledge is used as a bootstrap to infer automatically a segmentation of a word sequence into concepts using either integer linear optimisation or stochastic word alignment models (IBM models). The relation between automatically-derived and manually-defined task-dependent concepts is evaluated on a spoken dialogue task with a reference annotation.

**Keywords:** concept discovery ; segmental semantic annotation, language understanding ; latent Dirichlet analysis ; dialogue systems

## 1 Introduction

Spoken dialogue systems in the field of information query are basically used to interface a database with users using speech. When probabilistic models are used in such systems, good performance is at the price of collecting a large set of field data, which must be transcribed and annotated at the semantic level. These data allow then efficient models to be trained in a supervised way. However, the annotation process is costly and as a consequence represents a real difficulty hindering the widespread development of these systems. Therefore devising automatic methods for the annotation would be highly beneficial for portability to new tasks, domains or languages.

To give a full description of the architecture of a dialogue system is out of the scope of this paper. Instead we limit ourselves to briefly recall the main

words	concept	normalized value
donnez-moi	null	
le	refLink-coRef	singular
tarif	object	payment-amount-room
puisque	connectProp	imply
je voudrais	null	
une chambre	number-room	1
qui coûte	object	payment-amount-room
pas plus de	comparative-payment	less than
cinquante	payment-amount-integer-room	50
euros	payment-unit	euro

**Fig. 1.** Semantic concept representation for the query “give me the rate since I’d like a room charged not more than fifty euros”.

information pipeline: once a speech recognizer has transcribed the signal it is common (though avoidable for very simple tasks) to use a module dedicated to extract the meaning of the user’s queries. This meaning representation is then conveyed to an interaction manager that decides upon the next best action to perform considering the current user’s input and the dialogue history (defining the dialogue state all together). One of the very first steps to build the spoken language understanding (SLU) module is the identification of literal concepts in the word sequence hypothesised by the speech recogniser. An example of a semantic representation in terms of literal concepts is given in Figure 1. Once the concepts have been identified they can be further composed to form the overall meaning of the sentence, for instance by means of a tree representation based on hierarchical semantic frames [12].

To address the issue of concept tagging several techniques are available. Some of these techniques now classical rely on probabilistic models, that can be either discriminative or generative. Among these, the following models have been studied in the last decade: hidden Markov models, finite state transducers, maximum entropy Markov models, support vector machines, dynamic Bayesian networks and conditional random fields (CRF). In [6] it is shown that CRFs obtain the best performance on a tourist information retrieval task in French (MEDIA [2]), but also in two other comparable corpora in Italian and Polish.

To be able to apply any such technique, basic concept units have to be defined by an expert. In the best case, most of these concepts can be derived straightforwardly from the pieces of information lurking in the database tables (mainly table fields but not exclusively). Some others are general (dialogue units but also generic entities such as number, dates etc.). However, to provide efficient and usable information to the reasoning modules (the dialogue manager in our case) concepts have to be fine-grained enough and application-dependent (even general concepts might have to be tailored to peculiar uses). To that extent it seems out of reach to derive the concept definitions using a fully automatic procedure. Anyhow the process can be bootstrapped, for instance by induction

of semantic classes such as in [17] or [8]. Our assumption here is that the most time-consuming parts of concept inventory and data tagging could be obtained in an unsupervised way, even though a final (but hopefully minimal) manual procedure is still required to tag the classes so as to manually correct automatic annotation.

Unlike the previous attempts cited above which developed *ad hoc* approaches, we investigate here the use of broad-spectrum knowledge extraction methods. Topic as used in information retrieval systems is a notion related to that of concept in SLU. For a long time, the topic detection task was limited to associate a single topic with a document and thus was not fitted to the SLU task requirements. The recently proposed LDA technique allows a document to have a probabilistic representation as a mixture of topics. Then multiple topics can co-occur inside a document and the same topic can be repeated. From these characteristics it is possible to consider the application of LDA to unsupervised concept inventory and concept tagging for SLU. A shortcoming is that LDA does not take into account at all the sequentiality of the data. To address this issue we propose to conclude the procedure with a final step introducing specific constraints for a correct segmentation of the data: the assignments of topics proposed by LDA are modified to be more segmentally coherent. Heretwo variants are evaluated to derive an optimal concept segmentation using the probabilities output by the LDA inference step. First the recourse to a technique used in the Machine Translation field is considered: stochastic word alignment with IBM models. Then a more general framework for global optimisation under constraints is also investigated: integer linear programming.

The paper is organised as follows. Principles of automatic induction of semantic classes are presented in Section 2, followed by the presentation of an induction system based on LDA. The additional segmentation step is presented in Section 3 with two variants: stochastic word alignment with IBM models (GIZA) and integer linear programming (ILP). Then evaluations and results are reported in Section 4 on the French MEDIA dialogue task.

## 2 Automatic Induction of Semantic Classes

### 2.1 Context Modelling

The idea of automatic induction of semantic classes is based on the assumption that concepts often share the same context (syntactic or lexical). Implemented systems are based on the observation of co-occurring words according to two different ways. The observation of consecutive words (bigrams or trigrams) allows for the generation of lexical compounds supposed to follow syntactic rules. The comparison of right and left contexts considering pairs of words allows for the clustering of words and compound words into semantic classes.

In [17] and [14], iterative systems are presented. Their implementations differ in the metrics chosen to evaluate the similarity during the generation of syntactic rules and semantic classes, but also in the number of words taken into account

in a word context and the order of successive steps (which ones to generate first: syntactic rules or semantic classes?). An iterative procedure is executed to obtain a sufficient set of rules in order to automatically extract knowledge from the data.

While there may be still room for improvement in these techniques we decided to skip them and to focus instead on general knowledge extraction approaches in order to evaluate their potential. For that purpose a global strategy based on an unsupervised machine learning technique is adopted in our work to produce semantic classes.

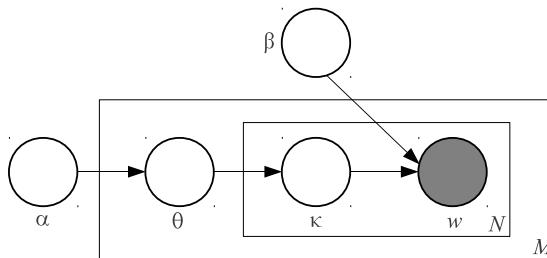
## 2.2 Implementation of an Automatic Induction System Based on Latent Dirichlet Allocation

Several approaches are available for topic detection in the context of knowledge extraction and information retrieval. They all more or less rely on the projection of the documents of interest in a semantic space to extract meaningful information. However, as the considered spaces (initial document words and latent semantics) are discrete performance of the proposed approaches for the topic extraction tasks are pretty unstable, and also greatly depend on the quantity of data available. In this work we were motivated by the recent development of a very attractive technique with major distinct features such as the detection of multiple topics in a single document. LDA [1] is the first principled description of a Dirichlet-based model of mixtures of latent variables. LDA will be used in our work to annotate the dialogue data in terms of topics in an unsupervised manner.

Basically LDA is a generative probabilistic model for text documents. LDA follows the assumption that a set of observations can be explained by latent variables. More specifically documents are represented by a mixture of topics (latent variables) and topics are characterized by distributions over words. The LDA parameters are  $\{\alpha, \beta\}$ .  $\alpha$  represents the Dirichlet parameters of  $K$  latent topic mixtures as  $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_K]$ .  $\beta$  is a matrix representing a multinomial distribution in the form of a conditional probability table  $\beta_{k,w} = P(w|k)$ , where  $k$  and  $w$  represent resp. a topic and a word. Based on this representation, LDA can estimate the probability of a new document  $d$  of  $N$  words  $d = [w_1, w_2, \dots, w_N]$  using the following procedure.

A topic mixture vector  $\theta$  is drawn from the Dirichlet distribution (with parameter  $\alpha$ ). The corresponding topic sequence  $\kappa = [k_1, k_2, \dots, k_N]$  is generated for the whole document accordingly to a multinomial distribution (with parameter  $\theta$ ). Finally each word is generated by the word-topic multinomial distribution (with parameter  $\beta$ , that is  $p(w_i|k_i, \beta)$ ). After this procedure, the joint probability of  $\theta$ ,  $\kappa$  and  $d$  is then:

$$p(\theta, \kappa, d|\alpha, \beta) = p(\theta|\alpha) \prod_{i=1}^N p(k_i|\theta)p(w_i|k_i, \beta) \quad (1)$$



**Fig. 2.** Graphical representation for LDA variables (from [1]). The grey circle is the only observable variable.

To obtain the marginal probability of  $d$ , a final integration over  $\theta$  and a summation over all possible topics considering a word is necessary:

$$p(d|\alpha, \beta) = \int p(\theta|\alpha) \left( \prod_{i=1}^N \sum_{k_i} p(k_i|\theta) p(w_i|k_i, \beta) \right) d\theta \quad (2)$$

The framework is comparable to that of probabilistic latent semantic analysis, but the topic multinomial distribution in LDA is assumed to be sampled from a Dirichlet prior and is not linked to training documents. This approach is illustrated in Figure 2.

Training of the  $\alpha$  and  $\beta$  parameters is possible using a corpus of  $M$  documents, with a fixed number  $K$  of topics to predict. A variational inference procedure is described in [1] which alleviates the intractability due to the coupling between  $\theta$  and  $\beta$  in the summation over the latent topics. Once the parameters for the Dirichlet and multinomial distributions are available, topic scores can be derived for any given document or word sequence.

In recent years, several studies have been carried out in language processing based on LDA. For instance, [18] worked on unsupervised language model adaptation; [4] ranked candidate passages in a question-answering system; [15] implemented LDA to classify short and sparse web texts.

In our work, LDA is employed to annotate each user's utterance of a dialogue corpus with topics. Utterances longer than one word are included in the training set as its sequence of *words*. Once the model has been trained, inference on data corpus assigns the topic with the highest probability to each word in a document. This probability is computed from the probability of the topic to appear in the document and the probability of the word to be generated by the topic. As a consequence we obtain a full topic annotation of the utterance.

Notice that LDA considers a user utterance as a bag of words. This implies that each topic is assigned to a word without any consideration for its immediate context. An additional segmental process is required if we want to introduce some context information in the topic assignment.

### 3 Segmental Annotation

#### 3.1 Benefits of a Segmental Annotation

The segmental annotation of the data is not a strict requirement for language understanding. Up to quite recently, most approaches for literal interpretation were limited to lexical-concept relations; for instance this is the case of the PHOENIX system [20] based on the detection of keywords. However, in an NLP perspective, the segmental approach allows the various levels of sentence analysis (lexical, syntactic and semantic) to be connected. Even though, in order to simplify its application, segments are generally designed specifically for the semantic annotation and do not have any constraint on their relation with the actual syntactic units (chunks, phrasal groups etc.). Ignoring such constraints not only simplifies the annotation process but also increases the robustness of the interpretation module, which will ultimately be used inside a spoken dialogue system and will have to deal with noisy data. In particular, highly spontaneous and ungrammatical utterances from the users, combined with errors from the speech recognizer are likely to severely decrease the performance of syntactic analysers.

Another interesting property of segmental approach is to offer a convenient way to dissociate the detection of a conceptual unit from the extraction of its associated value. The value corresponds to the normalisation of the surface form (see last column in Fig. 1); for instance if the segment “not more than” is associated to the concept *comparative-payment*, its value is “less than”. The same value would be associated with the word sequences “not exceeding” or “inferior to”. Value extraction requires therefore a link between concepts and words; the normalisation problem can be addressed by means of regular expressions or concept-dependent language models (even using integrated approaches such as described in [10]). In the case of global approaches (i.e. not segmental), value extraction must be done directly at the level of the conceptual unit tagging, as in [11]. This additional level is very complex (as some values may not be enumerable, such as numbers and dates) and is only affordable when the number of authorised values (for the enumerable cases) is low.

To refine the LDA output, the topic-to-word alignment proposed by LDA is discarded and an automatic procedure is used to derive the best alignment between topics and words. While the underlying probabilistic models are pretty comparable, the major interest of this approach is to separate the tasks of detecting topics and aligning topics with words. It is then possible to introduce additional constraints (such as locality, total number of segments, limits on repetitions etc.) in the latter task which would otherwise hinder topic detection. Conversely the alignment is self-coherent and able to question the associations proposed during topic detection with respect to its own constraints only. Two approaches were designed to this purpose: one based on IBM alignment models and another one based on integer linear optimisation.

### 3.2 Alignment with IBM Models (GIZA)

LDA is able to generate one topic  $k_i$  by word  $w_i$  from the probability  $p(k_i|w_i, d)$  that  $k_i$  is associated with  $w_i$  of the document  $d$ . However, this method has the drawback of considering a document as a bag-of-words and it ignores the effect of  $k_i$  on the choice of the topics of the surrounding words  $w_{i-1}$  and  $w_{i+1}$ . This independence assumption between consecutive topics tends to generate too many distinct topics for each utterance. To tackle this problem, we resorted to a statistical word-alignment method based on the so-called IBM models, widely used in Machine Translation [3].

The use of the IBM models in our context offers several interests. First, the output of LDA generates a topic sequence for each word sequence and can be thus seen as a translation pair of a word-topic parallel corpus. Contrary to LDA, the HMM model [19] as well as IBM models 2-4 take into account the word order in the translation pairs to align. Finally, the IBM models can be trained in an unsupervised way without providing any alignment between words and topics.

To refine the assignment of topics in the documents with the IBM models, we adopted the following approach. Since LDA tends to produce a too high number of different topics for each utterance, a filtering process is applied on the LDA outputs to keep only the most relevant topics for each document. We keep the  $\chi_{max}$  topics with the highest probabilities  $p(k_i|w_i, d)$  computed by LDA, where  $\chi_{max}$  is a value empirically fixed from the expected set of topics according to the number of words in an utterance. The IBM models are expected to correct some errors made by LDA, and in particular to assign words associated with the discarded topics to more likely topics. Two additional processing steps are applied on the topic sequences. Only the first occurrence is kept for consecutive repetitions of the same topic, as the IBM models can align several words with a single topic. Besides, the *null* concept (for out-of-domain or irrelevant word segments) is removed since this concept disturbs the alignments between words and concepts, as shown by a previous experiment done in a similar setup [7]. The IBM models already incorporate an empty word to account for source words that are not aligned with any target word; this empty word can be considered as the *null* concept for our task. This of course supposes that a manual tagging of LDA topics has allowed to define task-dependent concepts from topics (but can also be limited to identify the topics corresponding to the *null* concept).

In our experiments the statistical word alignment toolkit GIZA++ [13] is used to train the IBM models 1-4 as well as the HMM model. The IBM models are built using the expectation-maximization algorithm from a word-topic parallel corpus, where the topic sequences generated from LDA are disconnected from the words. To be able to train the most informative IBM model 4, the following training pipeline was considered: 5 iterations of IBM1, 5 iterations of HMM, 3 iterations of IBM3 and 3 iterations of IBM4. The IBM4 model obtained at the last iteration is finally used to align words and topics. In order to improve alignment, IBM models are usually trained in both directions (words towards concepts and *vice versa*) then symmetrised by combining them. For this purpose,



we resorted to the default symmetrization heuristics used in MOSES, a widely used machine translation system toolkit [9].

### 3.3 Alignment with Integer Linear Programming (ILP)

Another approach to the re-alignment of LDA outputs is based on a general optimisation technique. Integer Linear Programming (ILP) is a widely used tool for modelling and solving combinatorial optimisation problems. It broadly aims at formulating a decision process as a set of equations or inequations (called *constraints*) which are linear with regards to some *decision variables*. An ILP is also composed of a linear *objective function* which represents the criterion (for instance a cost, a profit or like in this study a probability). Solving an ILP consists in assigning values to decision variables, such that all constraints are satisfied and the objective function is optimized. Readers may refer to [5] for an overview of applications and methods of ILP.

We provide two ILP formulations for solving the topic assignment problem related to a given document. They both take as input data an ordered set  $d$  of words  $w_i$ ,  $i = 1 \dots N$ , a set of  $K$  available topics and, for each word  $w_i \in d$  and topic  $k = 1 \dots K$ , the natural logarithm of the probability  $p(k|w_i, d)$  that  $k$  is assigned to  $w_i$  in the considered document  $d$ . Model [ILP] simply finds the highest-probability assignment of one topic to each word in the document, such that at most  $\chi_{max}$  different topics are assigned.

$$[ILP] : \max \sum_{i=1}^N \sum_{k=1}^K \ln(p(k|w_i, d)) x_{ik} \quad (3)$$

$$\sum_{k=1}^K x_{ik} = 1 \quad i \quad (4)$$

$$y_k - x_{ik} \geq 0 \quad i, k \quad (5)$$

$$\sum_{k=1}^K y_k \leq \chi_{max} \quad (6)$$

$$x_{ik} \in \{0, 1\} \quad i, k$$

$$y_k \in \{0, 1\} \quad k$$

In this model, decision variable  $x_{ik}$  is equal to 1 if topic  $k$  is assigned to word  $w_i$ , and equal to 0 otherwise. Constraints 4 ensure that exactly one topic is assigned to each word. Decision variable  $y_k$  is equal to 1 if topic  $k$  is used. Constraints 5 force variable  $y_k$  to take a value of 1 if at least one variable  $x_{ik}$  is not null. Moreover, constraints 6 limit the total number of topics used. The objective function 3 merely states that we want to maximize the total probability of the assignment. Through this model, our assignment problem is identified as a *p-centre* problem (see [16] for a survey on such location problems).

Numerical experiments show that [ILP] tends to give sparse assignments: most of the time, adjacent words are assigned to different topics even if the total number of topics is correct. To prevent this unnatural behaviour, we modified [ILP] to consider groups of consecutive words instead of isolated words. Model

[*ILP\_seg*] partitions the document into segments of consecutive words, and assigns one topic to each segment, such that at most  $\chi_{max}$  segments are created. For the sake of convenience, we denote by  $\bar{p}(k|w_{ij}, d) = \sum_{l=i}^j \ln(p(k|w_l, d))$  the logarithm of the probability that topic  $k$  is assigned to all words from  $i$  to  $j$  in the current document.

$$[ILP\_seg] : \max \sum_{i=1}^N \sum_{j=i}^N \sum_{k=1}^K \bar{p}(k|w_{ij}, d) x_{ijk} \quad (7)$$

$$\sum_{j=1}^i \sum_{l=i}^N \sum_{k=1}^K x_{jlk} = 1 \quad i \quad (8)$$

$$\sum_{i=1}^N \sum_{j=i}^N \sum_{k=1}^K x_{ijk} \leq \chi_{max} \quad (9)$$

$$x_{ijk} \in \{0, 1\} \quad i, j, k$$

In this model, decision variable  $x_{ijk}$  is equal to 1 if topic  $k$  is assigned to all words from  $i$  to  $j$ , and 0 otherwise. Constraints 8 ensure that each word belongs to a segment that is assigned a topic. Constraints 9 limit the number of segments. Due to the small size of the instances considered in this paper, both [*ILP*] and [*ILP\_seg*] are well solved by a direct application of an ILP solver.

## 4 Evaluation and Results

### 4.1 MEDIA Corpus

The MEDIA corpus is used to evaluate the proposed approach and to compare the various configurations. MEDIA is a French corpus related to the domain of tourism information and hotel booking [2]. 1,257 dialogues were recorded from 250 speakers with a wizard of Oz technique (a human agent mimics an automatic system). This dataset contains 17k user utterances and 123,538 words, for a total of 2,470 distinct words.

The MEDIA data have been manually transcribed and semantically annotated. The semantic annotation uses 75 concepts (e.g. *location*, *hotel-state*, *time-month*...). Each concept is supported by a sequence of words, the *concept support*. The *null* concept is used to annotate every words segment that does not support any of the 74 other concepts (and does not bear any information wrt the task). On average, a concept support contains 2.1 words, 3.4 concepts are included in an utterance and 32% of utterances are restrained to a single word (generally “yes” or “no”). Table 1 gives the proportions of utterances according to the number of concepts in the utterance.

Notice that each utterance contains at least one concept (the *null* label being considered as a concept). As shown in Table 2, some concepts are supported by few segments. For example, 33 concepts are represented by less than 100 concept supports. Considering that, we can foresee that finding these poorly represented concepts will be extremely hard if not impossible for LDA.

**Table 1.** Proportion of user utterances as a function of the number of concepts in the utterance.

# concepts	1	2	3	[4,72]
% utterances	49.4	14.1	7.9	28.6

**Table 2.** Number of concepts according to their occurrence range.

# concept supports	[1,100[	[100,500[	[500,1k[	[1k,9k[	[9k,15k]
# concepts	33	21	6	14	1 ( <i>null</i> )

## 4.2 Evaluation Protocol

Unlike previous studies, we chose a fully automatic way to evaluate the systems. In [17], a manual process is introduced to reject induced classes or rules that are not relevant to the task and also to name the semantic classes with the appropriate label. Thus, they were able to evaluate their semi-supervised annotation on the ATIS corpus. In [14], the relevance of the generated semantic classes was manually evaluated giving a mark to each induced semantic rule.

To evaluate the unsupervised procedure it is necessary to associate each induced topic with a MEDIA concept. To that purpose, the reference annotation is used to align topics with MEDIA concepts at the word level. A co-occurrence matrix is computed and each topic is associated with its most co-occurring concept.

As MEDIA reference concepts are very fine-grained, we also define a *high-level* concept hierarchy containing 18 clusters of concepts. For example, a high-level concept *payment* is created from the 4 concepts *payment-meansOfPayment*, *payment-currency*, *payment-total-amount*, *payment-approx-amount*; a high-level concept *location* corresponds to 12 concepts (*location-country*, *location-district*, *location-street*, ...). Thus, two levels of concepts are considered for the evaluation: *high-level* and *fine-level*.

The evaluation is presented in terms of the classical F-measure, defined as a combination of precision and recall measures. Two levels are also considered to measure topic assignment quality:

- *alignment* corresponds to a full evaluation where each word is considered and associated with one topic/concept;
- *generation* corresponds to the set of topics/concepts generated for a turn (no order, no word-alignment).

## 4.3 System Descriptions

Four systems are evaluated in our experiments.

[LDA] is the result of the unsupervised learning of LDA models using GIBBS-LDA++ tool<sup>3</sup>. It assigns the most probable topic to each word occurrence in a

<sup>3</sup> <http://gibbslda.sourceforge.net/>

**Table 3.** Examples of topics discovered by LDA ( $K = 100$ ).

Topic 0 <i>information</i>		Topic 13 <i>time-date</i>		Topic 18 <i>sightseeing</i>		Topic 35 <i>politeness</i>		Topic 33 <i>location</i>		Topic 43 <i>answer-yes</i>	
words	prob	words	prob	words	prob	words	prob	words	prob	words	prob.
d'	0.28	du	0.16	de	0.30	au	0.31	de	0.30	oui	0.62
plus	0.17	au	0.11	la	0.24	revoir	0.27	Paris	0.12	et	0.02
infor-										abso-	
mations	0.16	quinze	0.08	tour	0.02	madame	0.09	la	0.06	lument	0.008
autres	0.10	dix-huit	0.07	vue	0.02	merci	0.08	près	0.06	autre	0.008
détails	0.03	décembre	0.06	Eiffel	0.02	bonne	0.01	proche	0.05	donc	0.007
obtenir	0.03	mars	0.06	sur	0.02	journée	0.01	Lyon	0.03	jour	0.005
										Notre-	
alors	0.01	dix-sept	0.04	mer	0.01	villes	0.004	aux	0.02	Dame	0.004
souhaite	0.003	nuits	0.04	sauna	0.01	bientôt	0.003	gare	0.02	d'accord	0.004

document as described in Section 2.2. This approach requires prior estimation of the number of clusters that are expected to be found in the data. To find an optimal number of clusters, we varied the number  $K$  of topics around 75, which is the number of reference concepts. 2k training iterations were made using default values for  $\alpha$  and  $\beta$ .

[*GIZA*] is the system based on the *GIZA++* toolkit<sup>4</sup> which re-aligns for each utterance the topic sequence assigned by [*LDA*] to word sequence as described in Section 3.2.

[*ILP*] and [*ILP\_seg*] systems are the results of the ILP solver IBM ILOG CPLEX<sup>5</sup> applied to the models described in Section 3.3.

For the three last systems, the  $\chi_{max}$  value has to be fixed according to the desired concept annotation. As on average a concept support contains 2.1 words,  $\chi_{max}$  is defined empirically according to the number of words in the sequence: if  $i \in [2, 4]$ :  $\chi_{max} = i$ ; if  $i \in [5, 10]$  words:  $\chi_{max} = i - 2$ ; and for utterances containing more than 10 words:  $\chi_{max} = i/2$ .

For the sake of simplicity, single-word utterances are processed separately with prior knowledge. City names, months, days or answers (e.g. “yes”, “no”, “yeah”) and numbers are identified in these one-word utterances.

#### 4.4 Generated Topic Observations

In Table 3, six topics generated by [*LDA*] are represented by their 8 highest probability words. For topic 13, it is interesting noting that words have quite similar weights. The most represented words are “du” (“from”) and “au” (“to”) and other words are numbers or months, which *a priori* corresponds to a “time-date” concept. For topic 43, the word “oui” (“yes”) is given a 0.62 probability,

<sup>4</sup> <http://code.google.com/p/giza-pp/>

<sup>5</sup> <http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/>

**Table 4.** Topic repartitions among the high-level or fine-level concepts for  $[LDA]$  and  $[ILP]$  ( $K = 100$ ).

		Topic 18 <i>sightseeing</i>		Topic 33 <i>location</i>		Topic 43 <i>answer-yes</i>	
		#occ	concept	$Ent(k)$	#occ	concept	$Ent(k)$
<b>[LDA]</b>							
high	292	Location		571	Location	705	Answer
	258	null	2.25	156	null	107	null
	94	Name		87	Comparative	27	Location
fine	258	null		226	loc.-distRel.	705	answer
	136	loc.-placeRel.	2.78	190	loc.-city	107	null
	100	loc.-distRel.		156	null	17	object
<b>[ILP]</b>							
high	300	Location		661	Location	846	Answer
	200	null	2.19	123	null	109	null
	102	Name		115	Comparative	24	Location
fine	200	null		234	loc.-distRel.	846	answer
	163	loc.-placeRel.	2.64	223	loc.-city	109	null
	98	name-hotel		129	loc.-placeRel.	16	name-hotel

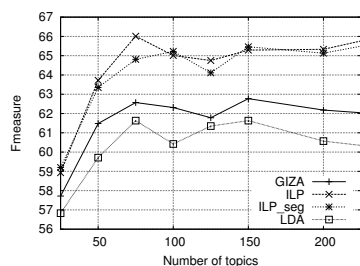
other words are “absolutely” or “okay” leading to an *a priori* “answer-yes” concept.

To observe which MEDIA concept is associated with these topics, the list of the 3 most co-occurring concepts and the number of co-occurrences are shown in Table 4. The 2 most co-occurring concepts in a topic are the same in  $[LDA]$  and  $[LDA + ILP]$ . However, the number of co-occurrences is higher in  $[ILP]$  than in  $[LDA]$ . An entropy measure  $Ent(k)$  is computed for each topic  $k$  in order to evaluate the reliability of the topic-concept association over all the possible concepts. It is computed as follows:

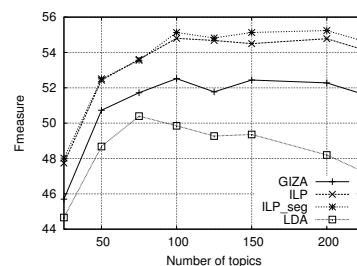
$$Ent(k) = - \sum_{\text{concepts } c} p(c|k) \log p(c|k) \quad (10)$$

with  $p(c|k) = \frac{\#(c \cap k)}{\#k}$

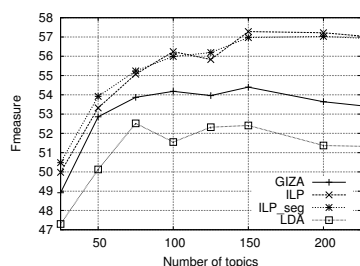
The topic entropy is always smaller considering  $[ILP]$  than  $[LDA]$ . This indicates that the re-assignment due to ILP alignment improves the reliability of the topic-concept association. Entropies measured with *high-level* concepts is always lower than with *fine-level* concepts, mainly because less classes are considered (18 instead of 75). Topic 18 is associated with the *Location* concept using the *high-level* configuration instead of the *null* one using the *fine-level* configuration, but the entropy is quite high. On the over hand, topic 43 has a low entropy, specifically with  $[ILP]$ . This shows that word “yes” is strongly associated with the concept “Answer”. Other topics representing the *null* concept can show entropies as low as 0.47 for topic 6 (“il”, “y”, “a”, “qu”, “est-ce”...).



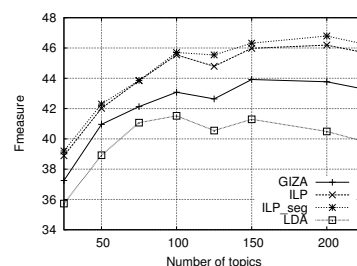
**Fig. 3.** F-measure of the high-level concept generation as a function of the number of topics.



**Fig. 4.** F-measure of the high-level concept alignment as a function of the number of topics.



**Fig. 5.** F-measure of the fine-level concept generation as a function of the number of topics.



**Fig. 6.** F-measure of the fine-level concept alignment as a function of the number of topics.

## 4.5 Results

Plots comparing the different systems implemented wrt the different evaluation levels in terms of F-measure are reported in Figures 3, 4, 5 and 6 (*high-level vs fine-level, alignment vs generation*).

The [LDA] system generates topics which are correctly correlated with the *high-level* concepts. It can be observed that the bag of 75 topics reaches an F-measure of 61.5% (Fig. 3). When not enough topics are required from [LDA], induced topics are too wide to fit the fine-grained concept annotation of MEDIA. On the other hand if too many high-level topics are required, the generation performance stays the same while a substantial decrease of the F-measure is observed in the *alignment* evaluation (Fig. 4). This effect can be explained by the automatic process to associate generated topics with reference concepts. Indeed, the increase of the number of topics makes them co-occur with many concepts, which often leads to assign them to the most frequent concept (*null*) in the studied corpus.

From the *high-level* to *fine-level* concept evaluations, results globally decrease by 10%. An additional global loss of 10% is also observed between the *generation* and *alignment* scoring methods. In the *fine-level* evaluation, a maximum F-measure of 52.2% is observed for the *generation* of 75 topics (Fig. 5) whereas the F-measure decreases to 41.5% in the *alignment* evaluation (Fig. 6).

To conclude on the [LDA] system, we can see that it generates topics having a good correlation with the *high-level* concepts, seemingly the best representation level between topics and concepts. From these results it seems obvious that an additional step is needed to obtain a more accurate segmental annotation, which is expected with the following systems.

The [GIZA] system improves the [LDA] results. It is very likely that the filtering process helps to discard the irrelevant topics. Therefore, the automatic alignment between words and the filtered topics induced by [LDA] with IBM models seems more robust when more topics (a higher value for  $K$ ) is required from [LDA], specifically in *high-level* concept *alignment* (Fig. 4).

Systems based on the ILP technique perform better than other systems whatever the evaluation. Considering [LDA] as the baseline, we can expect significant gains of performance. For example, an F-measure of 66% is observed for the ILP systems considering the *high-level* concept *generation* for 75 topics (Fig. 3), in which case the maximum for [LDA] was 61.5%, and an F-measure of 55% is observed (instead of 50.5% for [LDA]) considering the *high-level* concept *alignment* (Fig. 4).

No significant difference was finally measured between both ILP models for the concept generation evaluations. Even though [ILP\_seg] seems to obtain slightly better results in the *alignment* evaluation. This could be expected since [ILP\_seg] intrinsically yields alignments with grouped topics, closer to the reference alignment used for the evaluation.

It is worth noticing that unlike [LDA] system behaviour, the results of [ILP] are not affected when more topics are generated by [LDA]. A large number of topics enables [ILP] to pick up the best topic for a given segment in a longer selection list. As for [LDA], the same losses are observed between *high-level* and *fine-level* concepts and the *generation* and *alignment* paradigms. Nevertheless, an F-measure of 54.8% is observed with *high-level* concepts in the *alignment* evaluation (Fig. 4) that corresponds to a precision of 56.2% and a recall of 53.5%, which is not so low considering a fully-automatic high-level annotation system.

## 5 Conclusions and Perspectives

In this paper an approach has been presented for concept discovery and segmental semantic annotation of user's turns in an information-query dialogue system. An evaluation based on an automatic association between generated topics and expected concepts has shown that topics induced by LDA are close to *high-level* task-dependent concepts. The segmental annotation process increases performance both for the generation and alignment evaluations. On the whole these

results confirm the applicability of the technique to practical tasks with expected gains in data production.

Future work will investigate the use of n-grams to extend LDA and to increase its accuracy for providing better hypotheses to the following segmentation techniques. Besides, other levels of data representation will be examined (use of lemmas, *a priori* semantic classes like city names. . . ) in order to better generalise on the data.

## Acknowledgements

This work is supported by the ANR funded project PORT-MEDIA ([www.port-media.org](http://www.port-media.org)) and the LIA OptimNLP project ([www.lia.univ-avignon.fr](http://www.lia.univ-avignon.fr)).

## References

1. D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
2. H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa. Semantic annotation of the french MEDIA dialog corpus. In *Proceedings of the ISCA 9th European Conference on Speech Communication and Technology*, 2005.
3. P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.
4. A. Celikyilmaz, D. Hakkani-Tur, and G. Tur. LDA based similarity modeling for question answering. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, 2010.
5. D.-S. Chen, R. Batson, and Y. Dang. *Applied Integer Programming: Modeling and Solution*. Wiley, 2010.
6. S. Hahn, M. Dinarelli, C. Raymond, F. Lefèvre, P. Lehnen, R. De Mori, A. Moschitti, H. Hermann Ney, and G. Riccardi. Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech and Language Processing*, 19(6):1569–1583, 2010.
7. S. Huet and F. Lefèvre. Unsupervised alignment for segmental-based language understanding. In *Proceedings of the EMNLP 1st Workshop on Unsupervised Learning in NLP (UNSUP)*, 2011.
8. E. Iosif, A. Tegos, A. Pangos, E. Fosler-Lussier, and A. Potamianos. Unsupervised combination of metrics for semantic class induction. In *Proceedings of the IEEE/ACL Spoken Language Technology Workshop*, 2006.
9. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL, Companion Volume*, 2007.
10. F. Lefèvre. Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. In *Proceedings of the IEEE International Conference on Audio, Signal and Speech Processing*, 2007.
11. F. Mairesse, M. Gašić, F. Jurčiček, S. Keizer, B. Thomson, K. Yu, and S. Young. Spoken language understanding from unaligned data using discriminative classification models. In *Proceedings of the IEEE International Conference on Audio, Signal and Speech Processing*, 2009.



12. M.-J. Meurs, F. Lefèvre, and R. de Mori. Spoken language interpretation: On the use of dynamic bayesian networks for semantic composition. In *Proceedings of the IEEE International Conference on Audio, Signal and Speech Processing*, 2009.
13. F. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.
14. A. Pargellis, E. Fosler-Lussier, A. Potamianos, and C. Lee. Metrics for measuring domain independence of semantic classes. In *Proceedings of the 7th ISCA European Conference on Speech Communication and Technology*, 2001.
15. X. Phan, L. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceeding of the 17th ACM International Conference on World Wide Web*, 2008.
16. C.S. ReVelle and H.A. Eiselt. Location analysis: A synthesis and survey. *European Journal of Operational Research*, 165(1):1–19, 2005.
17. K. Siu and H. Meng. Semi-automatic acquisition of domain-specific semantic structures. In *Proceedings of the ISCA 6th European Conference on Speech Communication and Technology*, 1999.
18. Y. Tam and T. Schultz. Unsupervised language model adaptation using latent semantic marginals. In *Proceedings of ISCA INTERSPEECH*, 2006.
19. S. Vogel, H. Ney, and C. Tillmann. HMM-based word alignment in statistical translation. In *Proceedings of COLING*, 1996.
20. W. Ward. Understanding spontaneous speech. In *Proceedings of the IEEE International Conference on Audio, Signal and Speech Processing*, 1991.