CONCEPT DISCOVERY FOR LANGUAGE UNDERSTANDING IN AN INFORMATION-QUERY DIALOGUE SYSTEM

Nathalie Camelin, Boris Detienne, Stéphane Huet, Dominique Quadri and Fabrice Lefèvre
LIA - University of Avignon, BP 91228, 84911 Avignon Cedex 09, France

Keywords: Concept discovery, Language understanding, Latent Dirichlet analysis, Dialogue systems.

Abstract: Most recent efficient statistical approaches for natural language understanding require a segmental annotation of training data. Such an 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-steps alternative to the fully manual annotation of data: an initial unsupervised concept discovery, based on latent Dirichlet allocation, is followed by an automatic segmentation using integer linear optimisation. The relation between discovered topics and task-dependent concepts is evaluated on a spoken dialogue task for which a reference annotation is available. Topics and concepts are shown close enough to achieve a potential reduction of one half of the manual annotation cost.

1 INTRODUCTION

Generally, information-query spoken dialogue systems are used to interface a database with users orally. Once a speech recogniser has transcribed the signal, the meaning of the user's queries is extracted by a Spoken Language Understanding (SLU) module. The very first step of this module is the identification of literal concepts. These concepts are application-dependent and fine-grained so as to provide efficient and usable information to the following reasoning modules (e.g. the dialogue manager). To this respect they can also be composed in a global tree-based structure to form the overall meaning of the sentence.

To address the issue of concept tagging several techniques are available. Some of these techniques now classical rely on probabilistic models, which can be either discriminant or generative. To obtain good performance when probabilistic models are used in such systems, field data are to be collected, transcribed and annotated at the semantic level. It is then possible to train efficient models in a supervised manner. However, the annotation process is costly and constitutes a real hindrance to a widespread use of the systems. Therefore any means to avoid it would be highly appreciable.

It seems out of reach to derive the concept definitions from a fully automatic procedure. Anyhow the process can be bootstrapped, for instance by induction of semantic classes such as in (Siu and Meng, 1999) or (Iosif et al., 2006). 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 derived classes so as to manually correct the automatic annotation.

Unlike the previous attempts cited before, which developed ad-hoc approaches, in the work described here we investigate the use of broad-spectrum knowledge discovery techniques. In this context the notion most related to that of concept in SLU seems to be the topic, as used in information retrieval systems. For a long time, the topic detection task was limited to the association of a single topic to a document and thus did not fit our requirements. The recently proposed latent Dirichlet allocation (LDA) technique has the capacity to derive a probabilistic representation of a document as a mixture of topics. As such LDA can consider that several topics can co-occur inside a single document or sentence and that the same topic can be repeated.

From these favorable characteristics we consider the application of LDA to concept discovery for SLU. Anyhow, LDA does not take into account the sequentiality of the data (due to the *exchangeability* assumption). It is then necessary to introduce constraints for a better segmentation of the data: assignment of topics proposed by LDA is modified to be more coherent in a segmental way.

Camelin N., Detienne B., Huet S., Quadri D. and Lefevre F..
CONCEPT DISCOVERY FOR LANGUAGE UNDERSTANDING IN AN INFORMATION-QUERY DIALOGUE SYSTEM.
DOI: 10.5220/0003640500240029
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2011), pages 24-29
ISBN: 978-989-8425-79-9

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 and the additional step of segmentation using integer linear programming (ILP). Then evaluations and results are reported in Section 3 on the French MEDIA dialogue task.

2 AUTOMATIC INDUCTION OF SEMANTIC CLASSES AND ANNOTATION

2.1 Context Modelization

The main idea of automatic induction of semantic classes is based on the assumption that concepts often share the same context (syntactic or lexical) (Siu and Meng, 1999), (Pargellis et al., 2001). While there may be still room for improvement in these techniques we decided instead to investigate general knowledge discovery approaches in order to evaluate their potentiality.

In that purpose a global two-steps strategy is proposed: first semantic classes (topics) are induced; then topics are assigned to words with segmental constraints. The major interest of the 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, number and types of segments, etc) in the second task which would otherwise hinder topic detection in the first place.

Several approaches are available for topic detection in the context of knowledge discovery and information retrieval. In this work we were motivated by the recent development of a very attractive technique which has interesting distinct features such as the detection of multiple topics in a single document. LDA (Blei et al., 2003) is the first principled description of a Dirichlet-based model of mixtures of latent topic variables. It generates a set of topics with probabilities for each topic to be associated with a word in a sentence. In our case this knowledge is thereafter used to infer a segmentation of the sentence using integer linear optimisation.

2.2 Implementation of an Automatic Induction System based on LDA

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 $\{\alpha, \beta\}$. α represents the Dirichlet parameters of K latent topic mixtures as $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_K]$. β is a matrix representing a multinomial distribution in the form of a conditional probability table $\beta_{k,w} = P(w|k)$. Based on this representation, LDA can estimate the probability of a new document d of N words $d = [w_1, w_2, \ldots, w_N]$ using the following procedure.

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

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

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

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

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.

In recent years, several studies have been carried out in language processing using LDA. For instance, (Tam and Schultz, 2006) worked on unsupervised language model adaptation, (Celikyilmaz et al., 2010) ranked candidate passages in a question-answering system and (Phan et al., 2008) implemented LDA to classify short and sparse web texts.

We chose to use an implementation of LDA, GIBBSLDA++ tool,[1] to annotate each user's utterance of a dialogue corpus with topics. Each utterance of more than one word is included in the training set as its sequence of *words*. Once the models are trained, inference on data corpus assigns each word in a document with the highest probability topic. [*LDA*] system executes 2,000 iterations with the default parameters for α and β.

Notice that single-word utterances are processed separately using prior knowledge. Names of cities, month, day or short answers (*e.g.* "yes", "no", "yeah") and numbers are parsed in these utterances.

[1] http://gibbslda.sourceforge.net/

Indeed, LDA cannot be applied on such utterances where no co-occurrences can be observed.

2.3 Alignment with Integer Linear Programming (ILP)

The topic alignment problem we have to solve in this paper can be considered as a combinatorial optimization problem. ILP is a basic method to deal with such problems.

ILP is a mathematical method for determining the best way to optimize a given objective. More specifically, an integer linear model aims at optimizing a linear *objective function* (for instance representing a cost, a benefit, a probability...) subject to linear equalities or inequalities (named the *constraints*). Both the objective function and the constraints are expressed using integer unknown variables (called *decision variables*). The coefficients of the objective function and the constraints are the input data of the model. Consequently, solving an ILP consists in assigning integer values to decision variables, such that all constraints are satisfied and the objective function is optimized (maximized or minimized). We report the interested reader to (Chen et al., 2010) for an introduction on approaches and applications of ILP.

We propose an ILP formulation for solving the topic alignment problem for one document. The input data are: an ordered set d of words (indexed from 1 to N), a set of K available topics and, for each word $w_i \in d$ and topic $k = 1...K$, the natural logarithm of the probability $p(w_i|k)$ that k is assigned to w_i in the considered document. Model $[ILP]$ determines the highest-probability assignment of one topic to each word in the document, such that at most χ_{max} different topics are assigned.

$$\begin{array}{rll}
[ILP] : \max \sum_{i=1}^{N} \sum_{k=1}^{K} & p(w_i|k) x_{ik} & (3) \\
\sum_{k=1}^{K} x_{ik} = 1 & i = 1...N & (4) \\
y_k - x_{ik} \geq 0 & i = 1...N, k = 1...K & (5) \\
\sum_{k \in \kappa} y_k \leq \chi_{max} & k = 1...K & (6) \\
x_{ik} \in \{0,1\} & i = 1...N, k = 1...K & \\
y_k \in \{0,1\} & k = 1...K &
\end{array}$$

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 e.g. (ReVelle and Eiselt, 2005) for a survey on such location problems).

Since the number of instances considered in this work is small, [ILP] can be straightforwardly solved to optimality using a ILP solver as ILOG-CPLEX. Numerical results are reported in Section 3. The considered system is denoted $[LDA + ILP]$ since $p(w_i|k)$ are given by $[LDA]$. χ_{max} has been chosen according to the desired concept annotation. As on average a concept support contains 2.1 words, χ_{max} is defined empirically according to the number of words: with $i = [\![2,4]\!]$:$\chi_{max} = i$ with $i = [\![5,10]\!]$ words: $\chi_{max} = i - 2$ and for utterances containing more than 10 words: $\chi_{max} = i/2$.

3 EVALUATION AND RESULTS

3.1 MEDIA Corpus

The MEDIA corpus is used to evaluate the proposed approach. MEDIA is a French corpus related to the domain of tourism information and hotel reservation (Bonneau-Maynard et al., 2005). 1257 dialogues were recorded from 250 speakers with a WoZ technique (a human simulating an automatic phone server). In our experiments we only consider the 17k user utterances. This dataset contains 123,538 words, for a total of 2470 distinct words.

The MEDIA data have been manually transcribed and semantically annotated. The semantic annotation is rich of 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 word segment that does not support any of the 74 other concepts. Concepts do not appear at the same frequency as shown in Table 1. For example, 33 concepts (44% of the concepts) are supported by 100 occurences at most, while 15 concepts (21% of the concepts) present more than 1,000 occurences (only *null* is above 9,000).

Table 1: Number of concepts according to their occurrence range.

[1,100]	[100,500]	[500,1k]	[1k,9k]	[9k,15k]
33	21	6	14	1 (*null*)

On average, a concept support contains 2.1 words, 3.4 concepts are included in a turn and 32% of the utterances are single-word turns (generally *y*es or *n*o).

3.2 Automatic Evaluation Protocol

As MEDIA reference concepts are very fine-grained, we introduce a *high-level* concept hierarchy containing 18 clusters of concepts. For example, a *high-level* concept *payment* is created, corresponding to the four 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: the *high-level* (18 classes) and the *fine-level* (75 classes).

To evaluate the unsupervised procedure in a fully automatic way, it is necessary to associate each induced topic with a MEDIA concept. To that purpose, topics are aligned with concepts based on their word support for each utterance according to the reference annotation. A co-occurrence matrix is computed and each topic is associated to its most co-occurring concept. Table 2 analyses this automatic association for two values of K, the number of topics induced by LDA. Some concepts may not be associated with any topic, the *CC* column (Concept Coverage) gives the percentage of concepts that are associated with a topic. The *NNT* column (Not Null Topic) computes the percentage of topics not associated with the *null* concept.

Table 2: Analysis of the automatic association between topics and concepts. CC: concept coverage by the topics. NNT: % of topics not associated with *null*.

		K=50 topics		K=200 topics	
		CC	NNT	CC	NNT
high	LDA	61	60	72	51
	LDA+ILP	67	62	82	57
fine	LDA	21	50	34	47
	LDA+ILP	24	58	39	53

Considering [*LDA*] and *fine-level* concepts, only one fifth of the MEDIA concepts are retrieved for $K = 50$ and up to one third for $K = 200$. Though 72% of the *high-level* concepts are retrieved with $K = 200$. Considering $[ILP + LDA]$, this value even increases to 82%. This "lost concept" phenomenom at *high-level* can be explained by the fact that 72% of the concepts are supported by less than 500 concept supports, which seems a bit low for [*LDA*] to modelize them as a topic. $[LDA + ILP]$ helps to cover more concepts. Obviously, when more topics are induced, more concepts are covered. It is also interesting to notice that about half of the topics are associated with the *null* concept. When the number of topics is increased, more concepts are discovered but also more topics are associated with *null*.

3.3 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 month that *a priori* leads to a "time-date" topic. For topic 43, the word "oui" ("yes") is given a 0.62 probability, other words are "absolutely" or "okay" leading to an *a priori* "answer-yes" topic.

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

$$Ent(t) = -\sum_{concepts\ c} p(c|t)\log p(c|t) \quad (7)$$

$$with \qquad p(c|t) = \frac{\#(c \bigcap t)}{\# t}$$

The topic entropy is always smaller considering $[LDA + ILP]$ than [*LDA*]. This indicates that the reassignment due to ILP alignment improves the reliability of the topic-concept association. Entropies measured with *high-level* concepts are always lower than with *fine-level* concepts, in particular because less classes are considered (18 instead of 75). For topic 18, we can see that *high-level* enables to consider this topic as a *Location* concept and not a *null* one but the entropy is quite high. On the over hand, topic 43 shows a low entropy, specifically in $[LDA + ILP]$. This shows that word "yes" is strongly associated with concept "Answer". Other topics representing the *null* concept can show an entropy of 0.47 like the 6th topic ("there", "is", "what", "how", "does", ...)

3.4 Results

The evaluation is presented in terms of F-measure, combining precision and recall measures. Quality of topic assignment is considered also according to 2 levels:

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

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

Topic 0 *information*		Topic 13 *time-date*		Topic 18 *sightseeing*		Topic 35 *politeness*		Topic 33 *location*		Topic 43 *answer-yes*	
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
informations	0.16	quinze	0.08	tour	0.02	madame	0.09	la	0.06	absolument	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
alors	0.01	dix-sept	0.04	mer	0.01	villes	0.004	aux	0.02	Notre-Dame	0.004
souhaite	0.003	nuits	0.04	sauna	0.01	bientôt	0.003	gare	0.02	d'accord	0.004

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

		Topic 18 *sightseeing*			Topic 33 *location*			Topic 43 *answer-yes*		
		#occ.	concept	$Ent(t)$	#occ.	concept	$Ent(t)$	#occ.	concept	$Ent(t)$
LDA	high	292	Location		571	Location		705	Answer	
		258	null	2.25	156	null	1.72	107	null	1.10
		94	Name		87	Comparative		27	Location	
	fine	258	null		226	loc.-distanceRel.		705	answer	
		136	loc.-placeRel.	2.78	190	location-city	2.57	107	null	1.19
		100	loc.-distanceRel.		156	null		17	object	
LDA + ILP	high	300	Location		661	Location		846	Answer	
		200	null	2.19	123	null	1.52	109	null	0.76
		102	Name		115	Comparative		24	Location	
	fine	200	null		234	loc.-distanceRel.		846	answer	
		163	loc.-placeRel.	2.64	223	location-city	2.44	109	null	0.80
		98	name-hotel		129	loc.-placeRel.		16	name-hotel	

Plots comparing the different systems implemented w.r.t. the different evaluation levels in terms of F-measure are reported in Figures 1 and 2.

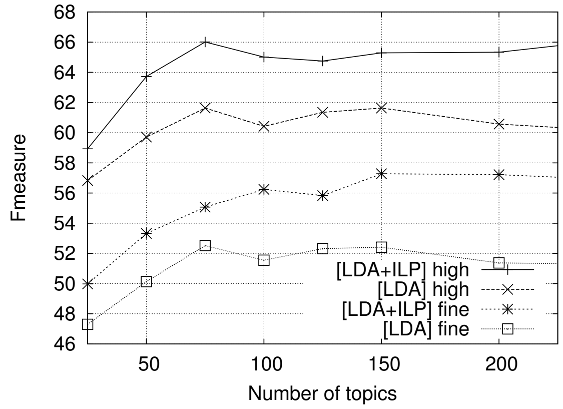

Figure 1: F-measure of the concept generation as a function of the number of topics.

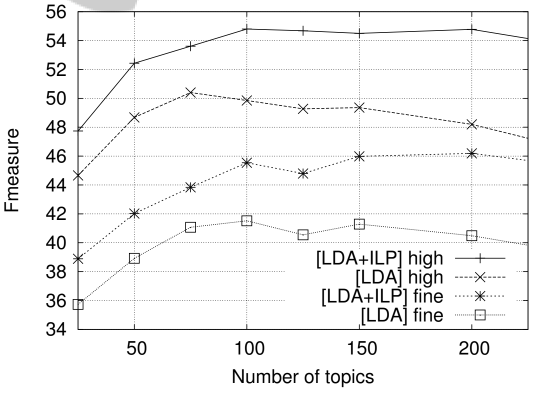

Figure 2: F-measure of the concept alignment as a function of the number of topics.

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 61.6% (Figure 1), corresponding to a precision of 59.4% and a recall of 64%. When $[LDA]$ is asked to generate too few topics, induced topics are not specific enough to fit the fine-grained concept annotation of MEDIA.

On the other hand, Figure 2 shows that a too high increase of the number of topics does not affect the bag of high-level topics significantly but induces a substantial decrease of the F-measure for the *alignment* evaluation. This effect can be explained by the automatic alignment method chosen to transpose topics into reference concepts. When there are too many topics, they co-occur with many concepts and are assigned to the most co-occurring one when some other concepts can co-occur only slightly less. In such situations, it is likely that *null* is the most co-occurring concept and the other concepts because they are too much scattered are not associated to enough topics. So they appear in the utterance but not on enough words to be retained by the segmentation process.

From the *high-level* to *fine-level* concept evaluation, results globally decrease of 10%. A loss of 12% is observed from the *generation* to the *alignment* evaluation. In the *fine-level* evaluation, a maximum F-measure of 52.5% is observed for the *generation* of 75 topics (Figure 1), corresponding to 54.9% in precision and 50.3% in recall whereas the F-measure decreases to 41% (precision=46.7% and recall=36.7%) in the *alignment* evaluation (Figure 2).

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. It is obvious that an additional step is needed to obtain a more accurate segmental annotation, what is expected with the use of ILP.

$[LDA+ILP]$ performs better whatever the level of evaluation. For instance, an F-measure of 66% is observed considering the *high-level* concept *generation* for 75 topics (Figure 2). As for $[LDA]$, the same losses are observed between *high-level* and *fine-level* concepts and *generation* and *alignment* paradigms. Nevertheless, an F-measure of 54.8% is observed at the *high-level* concept in *alignment* evaluation (Figure 2) 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.

4 CONCLUSIONS

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 been 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 gain 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. Also another technique for automatic re-alignment, based on IBM models used in stochastic machine translation, will be examined.

REFERENCES

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A., and Mostefa, D. (2005). Semantic annotation of the French MEDIA dialog corpus. In *Proceedings of Eurospeech*.

Celikyilmaz, A., Hakkani-Tur, D., and Tur, G. (2010). LDA based similarity modeling for question answering. In *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*, pages 1–9. Association for Computational Linguistics.

Chen, D., Batson, R. G., and Dang, Y. (2010). *Applied Integer Programming: Modeling and Solution*. Wiley.

Iosif, E., Tegos, A., Pangos, A., Fosler-Lussier, E., and Potamianos, A. (2006). Unsupervised combination of metrics for semantic class induction. In *Proceedings of the IEEE Spoken Language Technology Workshop*, pages 86–89.

Pargellis, A., Fosler-Lussier, E., Potamianos, A., and Lee, C. (2001). Metrics for measuring domain independence of semantic classes. In *Proceedings of Eurospeech*.

Phan, X., Nguyen, L., and Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceeding of the 17th international conference on World Wide Web*, pages 91–100. ACM.

ReVelle, C. S. and Eiselt, H. A. (2005). Location analysis: A synthesis and survey. *European Journal of Operational Research*, 165(1):1–19.

Siu, K. and Meng, H. (1999). Semi-automatic acquisition of domain-specific semantic structures. In *Proceedings of Eurospeech*.

Tam, Y. and Schultz, T. (2006). Unsupervised language model adaptation using latent semantic marginals. In *Proceedings of Interspeech*, pages 2206–2209.