Unsupervised Concept Annotation using Latent Dirichlet Allocation and Segmental Methods

Nathalie Camelin, Boris Detienne, Stéphane Huet, Dominique Quadri and Fabrice Lefèvre
LIA - University of Avignon, BP 91228
84911 Avignon Cedex 09, France
{nathalie.camelin,boris.detienne,stephane.huet,dominique.quadri,fabrice.lefevre}@univ-avignon.fr

Abstract

Training efficient statistical approaches for natural language understanding generally requires data with segmental semantic annotations. Unfortunately, building such resources is costly. In this paper, we propose an approach that produces annotations in an unsupervised way. The first step is an implementation of latent Dirichlet allocation that produces a set of topics with probabilities for each topic to be associated with a word in a sentence. This knowledge is then used as a bootstrap to infer a segmentation of a word sentence into topics using either integer linear optimisation or stochastic word alignment models (IBM models) to produce the final semantic annotation. The relation between automatically-derived topics and task-dependent concepts is evaluated on a spoken dialogue task with an available reference annotation.

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 can only be reached at the price of collecting a lot of field data, which must be transcribed and annotated at the semantic level. It becomes then possible to train efficient models in a supervised manner. However, the annotation process is costly and as a consequence represents a real difficulty hindering the widespread development of these systems. Therefore any means to avoid it would be profitable as portability to new tasks, domains or languages would be greatly facilitated.

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 that 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. 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 concept is given in Figure 1. Once the concepts are 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.

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 most efficiently studied this last decade are: hidden Markov models, finite state transducers, maximum entropy Markov models, support vector machines, dynamic fields (CRF). In (Hahn et al., 2010) it is shown that CRFs obtain the best performance on a tourist information retrieval task in French (MEDIA (Bonneau-Maynard et al., 2005)), but also in two other comparable corpora in Italian and Polish.

To be able to apply any such technique, basic con-

Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pages 72–81,
Edinburgh, Scotland, UK, July 27–31, 2011.

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

Figure 1: Semantic concept representation for the query *"give me the rate since I'd like a room charged not more than fifty euros"*.

cept 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 (dialogic units but also generic entities such as number, dates, etc). However, to provide an 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 (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 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. The notion most related to that of concept in SLU is the topic, as used in information retrieval systems. Anyhow for a long time, the topic detection task was limited to associate a single topic to a document and thus was not fitted to our requirements. The recently proposed LDA technique allows to have a probabilistic representation of a document 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 modelize at all the sequentiality of the data. To address this issue we propose to conclude the procedure with a final step to introduce specific constraints for a correct segmentation of the data: the assignments of topics proposed by LDA are modified to be more segmentally coherent.

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 step of segmentation is presented in Section 3 with two variants: stochastic word alignment (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 modeling

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) enables the generation of lexical compounds supposed to follow syntactic rules. The comparison of right and left contexts considering pairs of words enables to cluster words (and word compounds) into semantic classes.

In (Siu and Meng, 1999) and (Pargellis et al., 2001), 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 instead to investigate 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 LDA

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 the 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 (Blei et al., 2003) 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. Then the relation between automatic topics and expected concepts will be addressed manually.

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:

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

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

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

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 α and β parameters is possible using a corpus of documents, with a fixed number of topics to predict. A variational inference procedure is described in (Blei et al., 2003) which alleviates the intractability due to the coupling between θ and β 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, (Tam and Schultz, 2006) worked on unsupervised language model adaptation; (Celikyilmaz et al., 2010) ranked candidate passages in a question-answering system; (Phan et al., 2008) implemented LDA to classify short and sparse web texts.

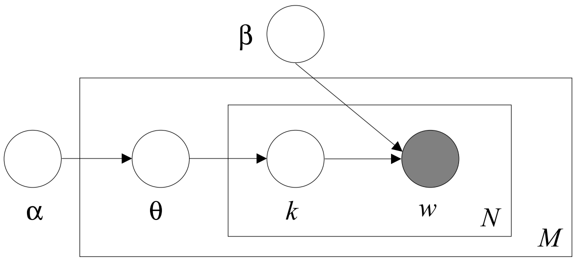

Figure 2: Graphical representation for LDA variables (from (Blei et al., 2003)). The grey circle is the only observable variable.

In our work LDA is employed to annotate each user's utterance of a dialogue corpus with topic. 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 (Ward, 1991) based on the detection of keywords. However in an NLP perspective, the segmental approach allows to connect the various levels of sentence analysis (lexical, syntactic and semantic). 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). To get relieved of such constraints not only simplifies the annotation process itself but as ultimately the interpretation module is to be used inside a spoken dialogue system, data will be noisy and generally bound the performance of the syntactic analysers (due to highly spontaneous and ungrammatical utterances from the users, combined with errors from the speech recognizer).

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 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 to "not exceeding" or "inferior to". Value extraction requires a link between concepts and words based on which the normalisation problem can be addressed by means of regular expressions or concept-dependent language models (even allowing integrated approaches such as described in (Lefèvre, 2007)). In the case of global approaches (not segmental), value extraction must be dealt with directly at the level of the conceptual unit tagging, as in (Mairesse et al., 2009). 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 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, 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)

Once topic assignments for the documents in the corpus have been proposed by LDA, a filtering process is done to keep only the most relevant topics of each document. The χ_{max} most probable topics are kept according to the probability $p(k|w_i, d)$ that topic k generated the word w_i of the document d. χ_{max} is a value fixed empirically according to the expected set of topics in a sentence. Then, the obtained topic sequences are disconnected from the words. At this point, the topic and word sequences can be considered as a translation pair to produce a word-topic parallel corpus. These data can be used with classical approaches in machine translation to align source and target sentences at the word level. Since these alignment models can align several words with a single topic, only the first occurrence is kept for consecutive repetitions of the same topic. These models are expected to correct some errors made by LDA, and to assign in particular words previously associated with discarded topics to more likely ones.

In our experiments the statistical word alignment toolkit GIZA++ (Och and Ney, 2003) is used to train the so-called IBM models 1-4 as well as the HMM model. 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*) and symmetrised by combining them. For this purpose, we resorted to the default symmetrization heuristics used by MOSES, a widely used machine translation system toolkit (Koehn et al., 2007).

3.3 Alignment with Integer Linear Programming (ILP)

Another approach to the re-alignment of LDA outputs is based on a general optimisation technique. ILP is a widely used tool for modelling and solving combinatorial optimisation problems. It broadly aims at modelling a decision process as a set of equations or inequations (called *constraints*) which are linear with regards to so-called *decision variables*. An ILP is also composed of a linear *objective function*. Solving an ILP consists in assigning values to decision variables, such that all constraints are satisfied and the objective function is optimised. We refer to (Chen et al., 2010) 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, $i = 1...N$, 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 $p(k|w_i, d)$ 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 χ_{max} different topics are assigned.

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

$$\sum_{k=1}^{K} x_{ik} = 1 \qquad i \tag{4}$$

$$y_k - x_{ik} \geq 0 \qquad i, k \tag{5}$$

$$\sum_{k=1}^{K} y_k \leq \chi_{max} \tag{6}$$

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

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

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 (ReVelle and Eiselt, 2005) 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 χ_{max} segments are created. For the sake of convenience, we denote by $\bar{p}(k|w_{ij}, d) = \sum_{l=i}^{j} log(p(k|w_l, d))$ the logarithm of the probability that topic k is assigned to all words from i to j in the current document.

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

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

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

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

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 (Bonneau-Maynard et al., 2005). 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 a utterance and 32% of the 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.

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

Table 1: Proportion of user utterances as a function of 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 hard for LDA.

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

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

4.2 Evaluation protocol

Unlike previous studies, we chose a fully automatic way to evaluate the systems. In (Siu and Meng, 1999), 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 (Pargellis et al., 2001), 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;
- *generation* corresponds to the set of topics 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 GIBBSLDA++ tool[1]. It assigns the most probable topic to each word occurrence in a 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 adjusted the number K of topics around the 75 reference concepts. 2k training iterations were made using default values for α and β.

$[GIZA]$ is the system based on the GIZA++ toolkit[2] which re-aligns for each sentence 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[3] applied to the models described in Section 3.3.

For the three last systems, the value χ_{max} has to be fixed 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: $\chi_{max} = i/2$.

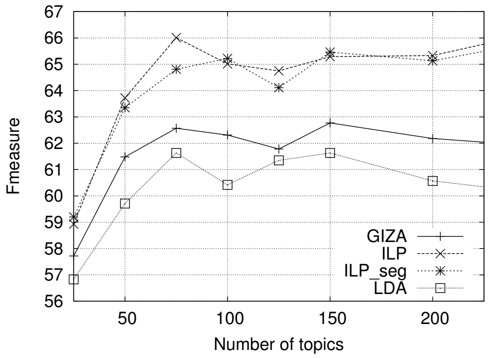

Figure 3: F-measure of the high-level concept generation as a function of the number of topics.

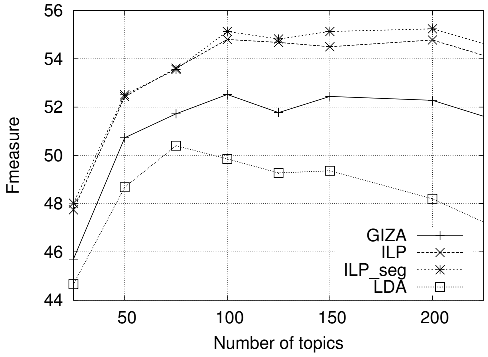

Figure 4: F-measure of the high-level concept alignment as a function of the number of topics.

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 Results

Examples of topics generated by $[LDA]$, with $K = 100$ topics, are shown in Table 3.

Plots comparing the different systems implemented w.r.t. 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

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

[2] http://code.google.com/p/giza-pp/

[3] http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/

Topic 0		Topic 13		Topic 18		Topic 35		Topic 33		Topic 43	
information		*time-date*		*sightseeing*		*politeness*		*location*		*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 3: Examples of topics discovered by LDA ($K = 100$).

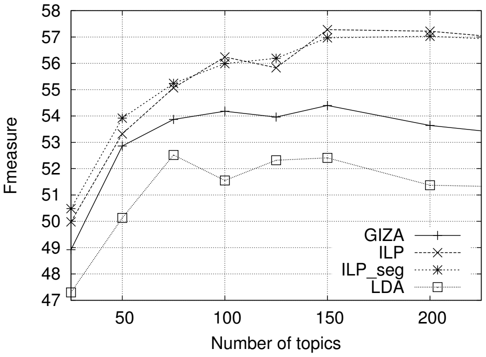

Figure 5: F-measure of the fine-level concept generation as a function of the number of topics.

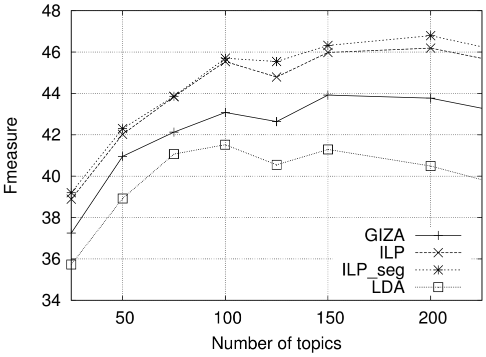

Figure 6: F-measure of the fine-level concept alignment as a function of the number of topics.

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 topics are required, the performance of bag of high-level topics 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 alignment method chosen to transpose topics into 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 for both the *generation* and *alignment* scorings. 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 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 (Figure 4), where 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*.

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 among in a longer selection list. 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 *alignement* evaluation (Figure 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 unsupervised approach for concept extraction and segmental annotation has been proposed and evaluated. Based on two steps (topic inventory and assignment with LDA, then re-segmentation with either IBM alignment models or ILP) the technique has been shown to offer performance above 50% for the retrieval of reference concepts. It confirms the applicability of the technique to practical tasks with an expected gain in data production.

Future work will investigate the use of n-grams to increase LDA accuracy to provide better hypotheses for the following segmentation method. 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) and the LIA OptimNLP project (www.lia.univ-avignon.fr).

References

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

H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa. 2005. Semantic annotation of the french media dialog corpus. In *Proceedings of the 9th European Conference on Speech Communication and Technology*.

A. Celikyilmaz, D. Hakkani-Tur, and G. Tur. 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.

Der-San Chen, Robert G. Batson, and Yu Dang. 2010. *Applied Integer Programming: Modeling and Solution*. Wiley, January.

Stefan Hahn, Marco Dinarelli, Christian Raymond, Fabrice Lefvre, Patrick Lehnen, Renato De Mori, Alessandro Moschitti, Hermann Ney, and Giuseppe Riccardi. 2010. Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech and Language Processing*, PP(99):1.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL, Companion Volume*, pages 177–180, Prague, Czech Republic.

F. Lefèvre. 2007. Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. In *Proceedings of ICASSP*, Honolulu, Hawai.

F. Mairesse, M. Gašić, F. Jurčíček, S. Keizer, B. Thomson, K. Yu, and S. Young. 2009. Spoken language understanding from unaligned data using discriminative classification models. In *Proceedings of ICASSP*, Taipei, Taiwan.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

A. Pargellis, E. Fosler-Lussier, A. Potamianos, and C.H. Lee. 2001. Metrics for measuring domain independence of semantic classes. In *Proceedings of the 7th European Conference on Speech Communication and Technology*.

X.H. Phan, L.M. Nguyen, and S. Horiguchi. 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.

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

K.C. Siu and H.M. Meng. 1999. Semi-automatic acquisition of domain-specific semantic structures. In *Proceedings of the 6th European Conference on Speech Communication and Technology*.

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

W Ward. 1991. Understanding Spontaneous Speech. In *Proceedings of ICASSP*, pages 365–368, Toronto, Canada.