Deliverable D4.2.3

Report on Topic Detection with Dirichlet Process Priors

Authors/Contributors: Stephen Pulman, stephen.pulman@comlab.ox.ac.uk, UOXF

Nigel Crook, nigel.crook@comlab.ox.ac.uk, UOXF

Ramon Granell, ramon.granell@comlab.ox.ac.uk, UOXF

Simon Dobnik, simon.dobnik@comlab.ox.ac.uk, UOXF

Manjari Chandran-Ramesh, manjari.chandran-ramesh@comlab.ox.ac

Work Package:	4.2
Lead Site:	UOXF
Due Date:	M48
Number of pages:	38
Deliverable Coordinator:	Nigel Crook, UOXF
Type:	Report and Software
Distribution:	External
Status:	Final

Area Coordinator:	Stephen Pulman (UOXF), Area 4
Project Coordinator:	Marc Cavazza, TEES
EU Project Officer:	Pierre-Paul Sondag
Contract No:	034434
Project Name:	Companions

Contents

ABSTRACT
This deliverable presents three software systems for topic detection. The first uses a natural language understanding approach, the second uses a Vector Space approach and the third uses a Dirichlet Process Mixture Model approach. This report describes each of these systems and presents experimental results obtained from them.

1 Introduction

In the HWYD scenario a human conversational partner describers their day at work and the Companion analyses lexical and emotional content of their utterances. If necessary, it asks the user clarification questions, and once a sufficient amount of information is gathered it generates an appropriate emotional response, both linguistically and in terms of the voice and body gestures. The conversations between a human and a HWYD Companion have two characteristics. The user can use free utterances which means that these may contain broken and elliptical sentences. Sometimes the utterances may also be quite long. They are collected from the Automatic Speech Recogniser (ASR) and as such they may be affected by recognition errors. The conversation with the ECA is not task oriented but is guided by the way a conversational partner introduces new facts. Below is a short excerpt from such conversation.

U: good evening sam (1)
S: hello steve how was your day (2)
U: not very good I had a computer crash so I can to work any more so what should I do (3)
S: who was responsible for this (4)
U: lighting technicians failed to properly maintain the computer (5)
S: what are you working on (6)
U: . . . (7)

Using standard natural language processing tools and models on such linguistic input presents a considerable challenge since the language models are typically built from written text. To extract the meaning of the user's utterance the HWYD Companion uses a Natural Language Understanding (NLU) component which is predominantly based on symbolic natural language processing techniques using hand crafted rules. The advantage of such approach is that we are able to extract linguistic predicate-argument dependencies at a considerable depth. However, because the linguistic input is sometimes incomplete and highly fragmented the rules cannot produce a good analysis each time. To improve the robustness of the system in such circumstances we have developed two alternative supplementary Topic Detection (TD) components which recognise the main topics or events in the user's utterances[1]. In this deliverable we describe all three of these approaches to topic detection.

2 A Symbolic approach to Topic Detection

The NLU processes the user's utterance(s) to create logical forms representing their meaning. At most levels of analysis the meaning is formalised as a set of objects consisting of feature-value pairs, some of which specify relations between the objects. The objects are passed to the Dialogue Manager (DM), the Emotional Model (EM), the Affective Strategy Model (ASM) and the Natural Language Generator (NLG).

In the HWYD prototype user utterances are recognised from speech to text by Nuance's Dragon NaturallySpeaking. The system returns multiple analyses but currently only the best one is used for further linguistic processing. The recognised utterance text is segmented and the segments are identified as instances of dialogue acts by Dialogue Act Tagger (DAT) [5]. The quality of the segmentation may effect further processing, for example in the fourth turn of the conversation above "lightning" was infelicitiously segmented as a part of the following sentence.

The NLU first tokenises the text from the dialogue act segments to a list of words and then applies it to a stochastic tagger using a Hidden Markov Model (HMM) that was trained on the Penn Treebank. For example, the user's utterance in line (3) of the example dialogue shown above would be tagged as:

> not/RB very/RB good/JJ I/PRP had/VBD a/DT computer/NN crash/NN so/IN I/PRP can/MD to/TO work/VB any/DT more/JJR so/IN what/WP should/MD I/PRP do/VB.

The tagged words are then grouped into units or chunks representing Noun Phrases (NP) and Verb Groups (VG) adapted from a method described in [10]. According to this method another HMM tagger is build on words but here the tags are B-X for a first word of a chunk of type X, I-X for a non-first word of a chunk of type X and O for a word not belonging to any chunk. Continuing with the previous example, this gives us analyses would give:

> not very good (NP I) (VP had) (NP a computer crash) so (NP I) (VP can to work) (NP any more) so (NP what) (VP should) (NP I) (VP do).

Semantically, NP and VG groups correspond to entities and events and as such they are the main units of the representation of the meaning of the utterance. In the NLU these are called objects. In the next step the internal structure of objects is determined using pre-defined rules based on the POS tags. For example, NPs are examined for a head noun and its determiners and modifiers. VGs are examined for a head verb, its

[1] In this context *event* means some natural event that has an emotional impact on the user.

adverbs, modal verbs, passive forms and negation particles. The components of objects are tested for grammatical features such as number and gender of NPs and number and tense of VGs. Finally, each object is classified as a Named Entity (NE) such as “person”, “event” and “emotion”. Again the process is based on rules which examine the context of the head word or match the headword or its modifiers to a NE class defined in the gazetteers. Some NE classes such as “person”, “location” and “temporal reference” are generic but most of those currently used had to be created specifically for the domain of conversations about office life. This includes, for example, events occurring at the office, parts of organisations and types of emotion. The information that is extracted from the internal structure of the NP and VG chunks is written as feature-value pairs of objects which become unification grammar categories. For example, one of the objects created from analysing the user’s utterance in line 3 of the example dialogue include:

```
[object(A:event),
attribute(A,nature,technical_problem),
attribute(A,type,crash),
attribute(A,gender,none),
attribute(A,number,sing),
attribute(A,determiner,a),
attribute(A,modifier,computer)].
```

At this stage of meaning analysis, objects created from NP and VG chunks only represent individual lexical items or concepts. During the next stage grammatical relations between them such as subject, object and different types of prepositional phrase modifiers are extracted from the sentences. These relations are also represented as feature-value pairs on objects. This is accomplished by a chart parser and set of phrase structure rules that can refer both to objects and the remaining tagged words. The ambiguity of parses is resolved with a simple heuristic which chooses the shortest path through the chart thus giving a preference to the deepest trees. The rules have been hand crafted and attempt to provide a general coverage of English. Because objects are also tagged with NE information, the rules can be lexicalised with domain-specific information from the HWYD scenario. For example, one could write a rule that applies to sentences such as “I met with Manjari about the Companions project” to identify “about NP:project” as the topic of the meeting event and “with NP:person” as participants of the meeting event. In practice, however, most lexical rules are implemented at a later stage of information extraction. The benefit of postponing this until then is that all lexical rules are kept in the same place.

The semantic representations obtained so far may contain anaphoric NPs such as pronouns which must be resolved to the correct referent. The system employs a reference resolution module based on [7] which has been adapted to the domain by giving more weight to the domain specific NE classes. The algorithm takes referring NPs (“Manjari”, “she”) as discourse referents and attempts to group them into co-reference classes by considering syntactic, semantic and salience properties of the utterance.

Overall, after completing the steps so far, representations such as the following are created.

```
[ A: event                               [ B: person
     agent = B,                               gender = male,
     nature = meet,                           is_user = true,
     temporal_reference = past,               number = sing,
     modal = null,                       ]
     with = C,
     about = E,
]

[ C: person                              [ E: project
     gender = female,                         type = project,
     name = Manjari,                          number = sing,
     number = sing,                           determiner = the,
     determiner = nulldet,                    modifier = Companions,
]                                        ]
```

These semantic representations represent the meaning of individual utterances. They do not yet correspond to the conceptual representations that other modules of the HWYD Companion, in particular the ASM, can work with. Such representations must follow the specifications of a commonly agreed ontology. Currently, the system can work with 95 events including sub-events or 51 events excluding sub-events. Events are the main conceptual categories corresponding to conversation topics. In addition there are 18 other object types corresponding to the NE classes. Typically, a set of semantic representation from the NL processing would correspond to one conceptual representation. This is because an event can be described in different ways, for example "I was made redundant", "I lost my job" and "I was fired". Both semantic and conceptual representations are given in the same formalism. This means that the semantic representations only need to be restructured to comply with the requirements of the ontology. This is accomplished by a set of hand crafted filtering rules. At a general level, the approach employs Information Extraction (IE) to dialogue analysis [6]. All three preceding sentences are represented with the following single analysis which is then sent to the DM.

```
[ A: event                               [ B: person
     agent = B,                               gender = male,
     nature = redundancy,                     is_user = true,
     temporal_reference = past,               number = sing,
]                                        ]
```

3 Vector Space Approach to Topic Detection

Vector Space Models [13, 2, 9] are widely used in the field of information retrieval. They constitute one of the fundamental methods for scoring documents when searching based on queries or for document clustering. In this method, each document d is represented as a vector, ($\overrightarrow{V_d}$) in a high dimensional space with each term in the document contributing to one of the dimensions. Hence the set of documents in the collection are represented as a set of vectors in the vector space. The advantages of vector space modelling is that they are mathematically sound and easily explained. However, they require each term in the vector to be given a weight. There are various methods that can be used to obtain this term weighting. One of the most popular methods is that of tf-idf weighting. The tf-idf weighting scheme is based on term frequency in the document and the inverse document frequency. We now define these quantities.

3.1 Term Frequency

The motivation behind term frequency is that a term that occurs more often in a document should have a higher score than terms that are less frequent. Hence each term in the document is given a weight that is proportional to the number of occurrences of that term in the document. This weighting scheme is referred to as ‘term frequency’ and is denoted as $tf_{t,d}$ where t is the term and d is the document. In this method, only the number of occurrences are considered and the exact ordering or the relative position of terms with other terms is not considered (i.e we treat a document as a “bag of words”). The aim of the representation is to score documents of similar content higher than documents that are not similar. It is not to capture semantic differences between documents on the same topic. Hence, only retaining information about the number of occurrences does not affect the final result. While the intuition that documents of similar content have similar words is reasonable, there are a number of words that are likely to occur in many documents. These terms then do not contribute towards discriminating between two documents, but have a high frequency of occurrence. This would then skew the similarity score between documents that have dissimilar content. Hence it is necessary to compensate for this within the weighting scheme so that all words are not equally important. Here the concept of inverse document frequency is introduced.

3.2 Inverse Document Frequency

The motivation behind inverse document frequency is to decrease the effect of terms that occur so often that they do not contribute towards relevance demarcation of documents. For this purpose the concept of document frequency is introduced. The document frequency, denoted as df_t, is defined as the number of documents in the collection that contain the term t. So this quantity is high for terms that commonly occur in the collection of documents (e.g. stop words) and low for words that do not occur so commonly (e.g. discriminating words). The inverse document frequency, denoted as idf_t is then defined as

$$idf_t = \log \frac{N}{df_t} \tag{1}$$

where N is the total number of documents in the collection. This makes the idf of an uncommon term high but commonly occurring terms have low idf values.

3.3 Tf-idf weighting

The tf-idf weighting is the combined weight of term frequency and inverse document frequency for each term in the document. It is given by

$$tf\_idf_{t,d} = tf_{t,d} \times idf_t \tag{2}$$

Multiplying the term frequency with the inverse document frequency causes the following:

- terms that occur in a small number of documents but with high frequency within each of those documents, namely, words having high relevance discriminating power, have high tf_idf values
- terms that occur in a large number of documents with less relevance discriminating power, have lower tf_idf values

3.4 Document Vectors and Similarity Measure

Using the tf_idf_t, the document vectors can now be obtained. If the term in the dictionary occurs in that document, the weight of that component is its tf_idf value. If the term in the dictionary does not occur in that document then the weight of that component is zero. Once the vectors have been obtained the documents can be compared using a similarity measure. The cosine similarity between two document vectors is a standard way of computing the relevance between the two documents. This is given by

$$sim(d_1, d_2) = \frac{\overrightarrow{V(d_1)} \cdot \overrightarrow{V(d_2)}}{|\overrightarrow{V(d_1)}||\overrightarrow{V(d_2)}|} \tag{3}$$

where the numerator is the dot product of the vectors of the two documents and the denominator is the product of the Euclidean length which compensates for the effect of the document length. In the case of a scoring if a query is relevant to a document, the similarity score is computed between the vectors of the query and the document.

3.5 Using Vector Space Modelling for Topic Modelling

The vector space is created by computing vectors for each document and the basis vectors are vectors of each term in the dictionary. In the vector space the term vectors are orthogonal. This set up is slightly modified to suit the application of vector space modelling to topic modelling as suggested by [3]. The basic premise here is each dimension of the space represents a fundamental topic and not a term in the dictionary. It is assumed that the topics are independent from each other and hence the vectors are all orthogonal to each other. The vectors of each topic consist of the weight for terms occurring in that topic and these weights represent the relevance of the term for that topic. Hence the weights are close to one for terms that are related to the topic and close to zero otherwise.

3.6 Obtaining Training Data

Before the topic of a query sentence can be identified using the vector space model, it is necessary to build the vector space. This was done using the British National Corpus. The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written. The Corpus was first stripped of all stop words. Then a list of keywords for each topic were created using WordNet, a lexical database for English. The BNC was then searched for occurrences of these keywords and a window of words on either side was extracted to build data for each topic. This window was considered as 10 words on either side so as to obtain a representative set of words for each topic.

3.7 Computing the vectors for the data

The vector space was built using the subset of data extracted from BNC and the algorithm given in Algorithm 1.

4 Dirichlet Process Mixture Models for Topic Detection

We compared the Vector Space model described in the previous section with a Dirichlet Process Mixture Model (DPMM) [8, 4, 1] for the unsupervised clustering of utterances into topic groups. This approach treats each utterance as a *bag of words* (i.e. an unordered collection of words) [14]. Utterances are clustered according to the relative counts of each word type that they contain so that utterances with similar histograms of word counts will, in general, appear in the same cluster. Dirichlet Processes offer one approach to developing Bayesian nonparametric mixture models. The remainder of this section introduces DPMMs, beginning with a brief look at finite Bayesian mixture models

```
/* Steps involved in identifying the topic of a user
utterance using Vector Space Model.  */
input  : keywords and window of size 10 words on either side for each topic obtained
         from BNC for training, user utterance
output: Topic of user utterance
begin
    foreach term in corpus do
        Count number of topics in which term occurs $df_t$.
        Calculate inverse document frequency where $N$ is the total number of topics
            $idf_t = \log \frac{N}{df_t}$
    end
    foreach topic in topic list do
        foreach term in topic do
            Calculate term frequency by counting number of occurrence
            Calculate tf-idf value for each term
                $tf\_idf_{t,d} = tf_{t,d} \times idf_t$
        end
        Build vector
    end
    foreach term in user utterance do
        Calculate tf-idf value
    end
    foreach topic in topic list do
        Calculate similarity measure between user utterance and topic
            $sim(UA, Topic_i) = \frac{\overrightarrow{UA} \cdot \overrightarrow{Topic_i}}{|\overrightarrow{UA}||\overrightarrow{Topic_i}|}$
    end
    Calculate maximum similarity measure. Corresponding topic is the topic of user
    utterance.
end
```

Algorithm 1: Topic Modelling using Vector Space Model

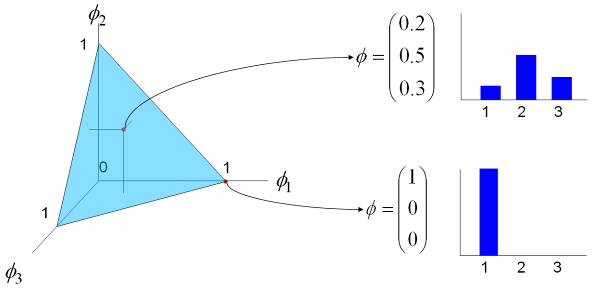
ϕ_2
1
$\phi = \begin{pmatrix} 0.2 \\ 0.5 \\ 0.3 \end{pmatrix}$
1 2 3
0
1
ϕ_1
1
$\phi = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$
ϕ_3
1 2 3

Figure 1: A 3-simplex with two examples points and the corresponding distributions

which will serve as useful background for presenting the Chinese Restaurant Process, the Dirichlet Process paradigm used in this work.

4.1 Finite Bayesian Mixture Models

A Dirichlet distribution is defined as a *measure on measures*. Specifically, a Dirichlet distribution defines a probability measure over the k-simplex. The k-simplex is a convex hull constructed so that each point on the surface of the simplex describes a probability distribution over k outcomes:

$$\begin{aligned} Q_k = \{(x_1, \ldots, x_k) : x_i \geq 0 & \\ & \forall i \in \{1 \ldots k\}, \sum_{i=1}^{k} x_i = 1\} \end{aligned}$$

Figure 1 shows a 3-simplex with two example points and the corresponding distributions. The Dirichlet distribution places a probability measure over the k-simplex so that certain subsets of points on the simplex (i.e. certain distributions) have higher probabilities than others (Figure 2). The probability measure in the Dirichlet is parameterised by a set of positive, non-zero concentration constants $\boldsymbol{\alpha} = \{\alpha_1, \ldots \alpha_k : \alpha_i > 0\}$, written $Dirichlet_k(\alpha_1, \ldots \alpha_k)$. The effects of different values of $\boldsymbol{\alpha}$ for the 3-simplex are shown in Figure 2.

The probability density function of the Dirichlet distribution is given by:

$$\begin{aligned} Dirichlet_k(\alpha_1, \ldots, \alpha_k) &= f(x_1, \ldots, x_k; \alpha_1, \ldots, \alpha_k) \\ &= \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{a_i - 1} \end{aligned}$$

Since a draw from a Dirichlet distribution (written $\beta \sim Dirichlet_k(\boldsymbol{\alpha})$) gives a distribution, a Dirichlet can be used as the prior for a Bayesian finite mixture model:

$$\beta \sim Dirichlet_k(\alpha_1, \ldots, \alpha_k)$$

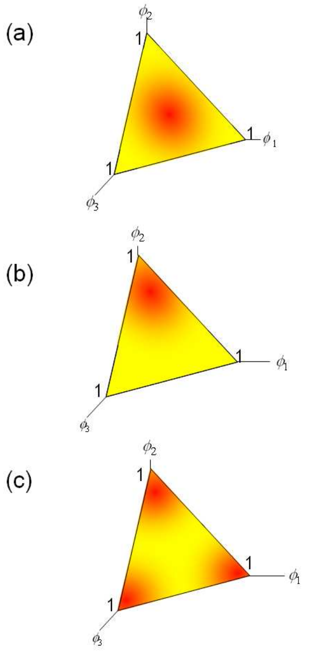

Figure 2: Three example Dirichlet Distributions over the 3-simplex. The darker areas show regions of high probability. The parameters for these distributions are (a) Dirichlet(5,5,5), (b) Dirichlet(0.2, 5, 0.2), (c) Dirichlet(0.5,0.5,0.5).

β is a distribution over the k components $\boldsymbol{\phi}$ of the finite mixture model. Each component ϕ_{z_i} is drawn from a base measure G_0 ($\phi_{z_i} \sim G_0$). The choice of distribution G_0 depends on the nature of the data to be clustered; with data that is represented using the *bag of words* model, G_0 must generate distributions over the word vocabulary. Hence the Dirichlet distribution is an appropriate choice in this case:

$$\phi_{z_i} \sim Dirichlet_v(\alpha_1, \ldots, \alpha_v)$$

where v is the size of the vocabulary.

For each data point (utterance) x_i a component ϕ_{z_i} is selected by a draw z_i from the multinomial distribution β:

$$z_i \sim Multinomial_k(\beta)$$

A suitable distribution $F(\phi_{z_i})$ is then used to draw the data point (utterance). In the bag of words model, the multinomial distribution is used to draw the words for each data point x_i:

$$x_i \sim Multinomial_v(\phi_{z_i})$$

A small example will illustrate this generative process. Imagine that there are just two types of utterances with a vocabulary consisting simply of the words A, B and C. A finite Bayesian mixture model in this case would first draw β from a suitable Dirichlet distribution (e.g. $\beta \sim Dirichlet_2(0.5, 1)$) as, for example, is shown in Figure 3(a). Next the two components ϕ_{z_1} and ϕ_{z_2} would be drawn from a suitable base distribution G_0 (e.g. $\phi_{z_1} \sim Dirichlet_3(1, 0.5, 0.5)$ and $\phi_{z_2} \sim Dirichlet_3(0.5, 0.5, 1)$, see Figure 3(b) and 3(c)). In this case, ϕ_{z_1} will tend to generate utterances containing more occurrences of word A than B or C, whilst ϕ_{z_2} will tend to generate utterances with more C's than A's or B's. A component z_i is then selected for each utterance ($z_i \sim Multinomial_k(\beta)$). Note that in this example, the distribution β would lead to more utterances generated by ϕ_{z_2} than by ϕ_{z_1}. Suppose that five utterances are to be generated by this model and that the components for each utterance are $z_1 = 1$, $z_2 = 2$, $z_3 = 2$, $z_4 = 1$ and $z_5 = 2$. The words in each utterance are then generated by repeated draws from the corresponding component (e.g. $x_1 = ACAAB$, $x_2 = ACCBCC$, $x_3 = CCC$, $x_4 = CABAAC$ and $x_5 = ACC$).

4.2 Dirichlet Processes

A Dirichlet Process can be thought of as an extension of a Dirichlet distribution where the dimensions of the distribution are infinite. The problem with the infinite dimension Dirichlet distribution, though, is that its probability mass would be distributed across the whole of the distribution. However, in most practical applications of mixture modelling there will be a finite number of clusters. The solution is to have a process which will tend to place most of the probability mass at the beginning of the infinite distribution, thereby making it possible to assign probabilities to clusters without restricting the number of clusters available. The GEM *stick breaking* construction (the name comes from the first

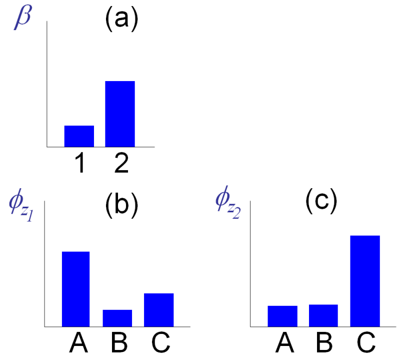

Figure 3: An example finite Bayesian mixture model. (a) The prior distribution over components ϕ_{z_1} (b) and ϕ_{z_2} (c)

letters of Griffiths, Engen and McCloskey [11]) achieves precisely this [12]. Starting with a stick of unit length, random portions β'_k are repeatedly broken off the stick, with each part that is broken off representing the proportion of probability assigned to a component:

$$\beta'_k \sim Beta(1, \alpha)$$
$$\beta_k = \prod_{i+1}^{k-1} (1 - \beta'_i) \cdot \beta'_k$$

The Dirichlet Process mixture model can now be specified as:

$$\beta \sim GEM(\boldsymbol{\alpha})$$
$$\phi_{z_i} \sim G_0 \quad z_i \in (1 \ldots \infty)$$
$$z_i \sim Multinomial(\beta)$$
$$x_i \sim F(\phi_{z_i})$$

4.3 Chinese Restaurant Process

The Chinese Restaurant Process (CRP) is a popular Dirichlet Process paradigm that has been successfully applied to many clustering problems. In the CRP, one is asked to imagine a Chinese restaurant with an infinite number of tables. The customers enter the restaurant and select, according to a given distribution, a table at which to sit. All the customers on the same table share the same dish. In this paradigm, the tables represent data clusters, the customers represent data points (x_i) and the dishes represent components (ϕ_z). As each customer (data point) enters the restaurant the choice of which table (cluster) and therefore which dish (component) is determined by a draw

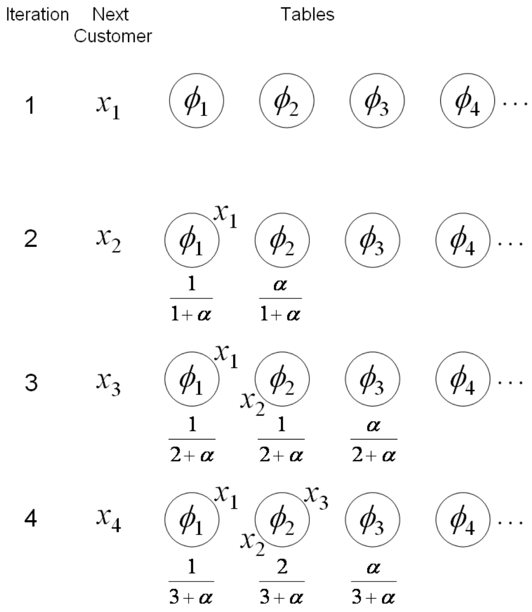

Figure 4: The first four steps of the initial clustering process of the CRP. The probability distribution over the tables is also shown in each case.

from the following distribution:

$$\phi_i|\phi_1,\ldots,\phi_{i-1} \sim \frac{1}{(\alpha+i-1)}\left(\sum_{j=1}^{i-1}\delta_{\phi_j}+\alpha G_0\right)$$

where α is the concentration parameter for the CRP. The summation over the δ_{ϕ_j}'s counts the number of customers sat at each of the occupied tables. The probability of sitting at an already occupied table, therefore, is proportional to the number of customers already sat at the table, whilst the probability of starting a new table is proportional to αG_0. Figure 4 illustrates four iterations of this initial clustering process.

Once all the customers (data points) have been placed at tables (clusters), the inference process begins. The posterior $p(\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{z}|\boldsymbol{x})$ cannot be calculated exactly, but Gibbs sampling can be used. Gibbs sampling for the CRP involves iteratively removing a randomly selected customer from their table, calculating the posterior probability distribution across all the occupied tables together with a potential new table (with a randomly drawn dish, i.e. component), and making a draw from that distribution to determine the new table for that customer. The posterior distribution across the tables is calculated as follows:

$$\begin{aligned} &\phi_i|\phi_1, \ldots, \phi_{i-1}, \boldsymbol{x} \\ &\qquad \sim \frac{1}{B}\left(\sum_{j=1}^{i-1} \delta_{\phi_j} p(x_i|\phi_j) + \alpha G_0 p(x_i|\phi_i)\right) \\ &B = \alpha p(x_k) + \sum_{j-1}^{i-1} p(x_i|\phi_i) \end{aligned}$$

In our experiments, we found that the following non-Bayesian procedure for updating the table components ϕ_i performed significantly better than the CRP process. After a predetermined number of samples, the dish (component) of each occupied table is updated to further resemble the customers (data points) sitting around it. In the *bag of words* approach used here, this involves converting the histogram of word counts in each customer (utterance) sitting at the table into an empirical distribution $\mathcal{H}(x_i)$, taking the average of these empirical distributions and modifying the dish (component) to further resemble this distribution:

$$\phi_i = \phi_i + \frac{\mu}{m_i} \sum_{j=1}^{m_i} \mathcal{H}(x_j)$$

where μ $(0 \leq \mu < 1)$ is the learning constant and m_i is the number of customers around table i. The inference process continues to iterate between Gibbs sampling and updating the table dishes (components) until the process converges. Convergence can be estimated by observing n consecutive samples in which the customer was returned to the same table they were taken from.

5 Experimental Results

Both the Vector Space model and the Chinese Restaurant Process model were trained using extracts from the BNC. These extracts were chosen according to a list of keywords each of which corresponded to one of the topics that can currently be recognised by the COMPANIONS English demonstrator. A subset of these keywords is shown below:

accommodation	bonus	break
deadline	deliverable	demotion
holiday	interview	management
meeting	move	occasion
perk	politics	presentation
project	promotion	report
review	technical	training
travel	weather	workload

Table 1: Example user utterances with the output of the TD

User Utterance	TD Output
I was late for the deadline	deadline
I missed the deadline	deadline
we met for coffee today	break
I gave a talk to the department	presentation
we had to submit the report by Christmas	deadline
I am getting a pay rise	review
I hope to get a promotion	promotion
I am working on a project	deliverable
I wrote several sections of the report	report
David is going on holiday next week	break

The procedure for taking the extracts from the BNC was as follows. First, upto 200 occurrences of each keyword were identified in the BNC. [2] Then a window of 21 non-stop words (10 on either side of the keyword + the keyword itself) were extracted to form the training set.

5.1 Vector Space Model Experiments and Results

A small prototype vector space model was trained using the BNC extracts for just 10 topic keywords:

bonus	break
deadline	deliverable
marriage	presentation
project	promotion
report	review

Table 1 shows the output from the trained small VSM model for several sentences relevant to the HWYD scenario. It is worth noting that in some cases the correct topic is identified even when the topic keyword does not appear in the utterances. For example, “we met for coffee today” is correctly classified as “break”, and “we had to submit the report by Christmas” is correctly classified as a ‘deadline’. In one or two cases the classification is appropriate, but not the best possible classification for those utterances; “I am working on a project” would be better classified as ‘project’ rather than ‘deliverable’, although ‘deliverable’ could still be relevant to this utterance.

[2] There were less than 200 occurrences in the BNC of some of the keywords, in which case the maximum number of occurrences were used.

This small VSM model was also evaluated on 49 user utterances that were spoken and logged during several user conversations with the full COMPANIONS English demonstrator. The utterances were chosen so that each of them was relevant to at least one of the 10 topics that the model was trained on. The ASR output for each utterance was presented as input to the topic detection model. The outputs of the model are shown in the table below. Here it can be seen that the model correctly classifies 22 out of the 49 utterances (45% correct). It should be noted that the cases where the model did not give the most relevant classification were categorized as incorrect in this table. The classification of 'deliverable' for utterance 18, for example, is not unreasonable but 'deadline' would have been better. Consequently, the 45% accuracy assessment is a little harsh. It should also be noted that the ASR output is poor in places resulting in a number of speech recognition errors (e.g. in utterances 5 and 17).

The small VSM model's responses to ASR input.

	User Utterance	**TD Output**	**Correct Topic**
1	well there 's a possibility that I might get a promotion	promotion	Y
2	yes it 's difficult to have to wait until you know the promotion	promotion	Y
3	before that I wouldn't go bark either my promotion or a possible pay rise	promotion	Y
4	on working under companions project but now I need a holiday is	deadline	N
5	so with any luck ahhh chief finish my presentation before the deadline	deadline	Y
6	the deadlines in the companions project	deliverable	N
7	and the deadline is next week	deadline	Y
8	is an important presentation because my boss will be there	presentation	Y
9	and my performance review is coming up	break1	N
10	if the performance review is successful I'm hoping I'll get a pay rise	deadline	N
11	but it all depends on the presentation being successful	presentation	Y
12	will issue gets a reputation for being difficult that that will not be good for her career	promotion	Y
13	Barbara is working on a new project	deadline	N
14	and I hope I will get a new computer in time to finish the presentation before the deadline	deadline	Y

15	I think I may be late meeting my deadline	deliverable	N
16	if I don't meet the deadline my boss will be very angry with me	deadline	Y
17	uhhh that Sir uhhh party are said not to project	marriage	N
18	I think I may be late meeting my deadline	deliverable	N
19	if I don't meet the deadline my boss will be very angry with me	deadline	Y
20	and that 's a not a good thing either performance review coming up next week	break1	N
21	I have a performance review next week and I had to prepare a presentation	deadline	N
22	if the performance review is successful I might get promotion	promotion	Y
23	as I started working on a presentation on my computer crashed my laptop is completely broken	marriage	N
24	then I can't make my performance review successful	bonus	N
25	then I will not get a promotion until the	promotion	Y
26	I think you got things the wrong way round you should be happy that I'm getting promoted but I don't think I am getting promoted	deliverable	N
27	the meeting was about a new project to uhhh going to start	deliverable	N
28	yet I need lots of coughing so that I can work on my presentations	report	N
29	I'm working on a presentation to my performance review	bonus	N
30	all as long as I get the presentation ready to the deadline next week	deadline	Y
31	that will be difficult in every way Matilda I won't be ahhh to get the presentation finished	presentation	Y
32	and I hope that my computer when crashed another presentation will be successful	marriage	N
33	if the presentation is good I hope I get a promotion	promotion	Y
34	while in these situations are a good presentation usually leads to promotion and even a pay rise	presentation	Y
35	so I got a good start working on my presentation next week	break1	N

36	so I got a good start on the presentation were then the Saturn in my laptop crashed	marriage	N
37	if I don't get a replacement computer ahhh be late with the deadline	deadline	Y
38	the deadline so the companions project	deliverable	N
39	the person who leaves the companions project is Mark	presentation	N
40	the presentation is important because it will form part of my performance review	bonus	N
41	it might be it depends how well the presentation goes my boss could be very angry with me	review	N
42	but if the presentation goes well I could get promotion	promotion	Y
43	yet promotion would be a good thing I hope I get a pay rise to	promotion	Y
44	it will be difficult we have a deadline coming up with a lot of work to do	deadline	Y
45	and Deborah were supposed to produce a presentation in	deadline	N
46	well were all working at the moment the deadline is next week	deadline	Y
47	it 's in the companions project	deliverable	N
48	the manager of the companions project is Mark	presentation	N
49	he tells us all what to do was part of the project plan	break	N

A large VSM model was trained on extracts from the BNC corpus using the following 24 topic keywords:

accommodation
bonus
break
deadline
deliverable
demotion
holiday
interview
management
meeting
move
occasion
perk
politics
presentation
project
promotion
report
review
technical
training
travel
weather
workload

The following table shows the output from the large VSM for the same ASR input that was presented to the small VSM model. The table shows a significant drop in performance for the topic detection task: The large VSM model only correctly classifies 9 out of the 49 user utterances (18% correct). The most likely explanation for this drop in performance is the significant increase in the dimensions of the vector space: The vector space for the small model was 10,872 whereas for the large model this increased to 112,182. A much larger number of training examples would be needed to successfully train a vector space model with these dimensions.

The large VSM model's responses to ASR input

	User Utterance	TD Output	Correct Topic
1	well there 's a possibility that I might get a promotion	move	N
2	yes it 's difficult to have to wait until you know the promotion	project	N
3	before that I wouldn't go bark either my promotion or a possible pay rise	weather	N
4	on working under companions project but now I need a holiday is	politics	N
5	so with any luck ahhh chief finish my presentation before the deadline	report	N
6	the deadlines in the companions project	promotion	N
7	and the deadline is next week	deadline	Y
8	is an important presentation because my boss will be there	promotion	N
9	and my performance review is coming up	deadline	N
10	if the performance review is successful I'm hoping I'll get a pay rise	occasion	N
11	but it all depends on the presentation being successful	management	N
12	will issue gets a reputation for being difficult that that will not be good for her career	project	N
13	Barbara is working on a new project	holiday	N
14	and I hope I will get a new computer in time to finish the presentation before the deadline	project	N
15	I think I may be late meeting my deadline	deadline	Y
16	if I don't meet the deadline my boss will be very angry with me	interview	N
17	uhhh that Sir uhhh party are said not to project	management	N
18	I think I may be late meeting my deadline	deadline	Y
19	if I don't meet the deadline my boss will be very angry with me	interview	N
20	and that 's a not a good thing either performance review coming up next week	weather	N
21	I have a performance review next week and I had to prepare a presentation	project	N

22	if the performance review is successful I might get promotion	occasion	N
23	as I started working on a presentation on my computer crashed my laptop is completely broken	presentation	Y
24	then I can't make my performance review successful	training	N
25	then I will not get a promotion until the	promotion	Y
26	I think you got things the wrong way round you should be happy that I'm getting promoted but I don't think I am getting promoted	promotion	Y
27	the meeting was about a new project to uhhh going to start	holiday	N
28	yet I need lots of coughing so that I can work on my presentations	move	N
29	I'm working on a presentation to my performance review	project	N
30	all as long as I get the presentation ready to the deadline next week	project	N
31	that will be difficult in every way Matilda I won't be ahhh to get the presentation finished	weather	N
32	and I hope that my computer when crashed another presentation will be successful	deadline	N
33	if the presentation is good I hope I get a promotion	occasion	N
34	while in these situations are a good presentation usually leads to promotion and even a pay rise	management	N
35	so I got a good start working on my presentation next week	occasion	N
36	so I got a good start on the presentation were then the Saturn in my laptop crashed	presentation	Y
37	if I don't get a replacement computer ahhh be late with the deadline	occasion	N
38	the deadline so the companions project	promotion	N
39	the person who leaves the companions project is Mark	promotion	N
40	the presentation is important because it will form part of my performance review	move	N
41	it might be it depends how well the presentation goes my boss could be very angry with me	management	Y

42	but if the presentation goes well I could get promotion	politics	N
43	yet promotion would be a good thing I hope I get a pay rise to	weather	Y
44	it will be difficult we have a deadline coming up with a lot of work to do	move	N
45	and Deborah were supposed to produce a presentation in	project	N
46	well were all working at the moment the deadline is next week	meeting	N
47	it 's in the companions project	management	N
48	the manager of the companions project is Mark	promotion	N
49	he tells us all what to do was part of the project plan	occasion	N

A major limitation of both the small and large models is that they were trained on extracts from the BNC, which covers a much broader spectrum of topics than those required by the HWYD scenario. Ideally, the models should have been trained on the HWRD corpus to ensure the creation of an appropriate set of vectors for representing topics in the HWYD domain. But this became available too late in the project for it to be deployed in this work.

5.2 CRP Experiments and Results

The topic extracts from the BNC were also clustered using the Chinese Restaurant Process described in Section 4.3. Up to 200 samples were used for each of the 24 topics. The full set of clustering results are shown in Appendix A. Table 2 summarises the distribution of the BNC extracts for each topic across the clusters. The results show that the CRP has been partially successful in clustering the topics solely based on the *bag of words* used in each topic window. Several topics have large numbers of extracts clustered in cluster 0. Tables 3 and 4 show the clusters that have 60% or more members from the same topic. These clusters could be seen as specialising on these topics. This deliverable is accompanied by a CD containing the Chinese Restaurant Process software that produced these results. The CD includes a README.txt file which explains how to configure and run the software.

6 Conclusion

This deliverable has presented work that has been done in developing the Topic Detector (TD) module for the COMPANIONS English demonstrator. The symbolic approach to topic detection undertaken by the NLU has be outlined. Two non-symbolic approaches to topic detection have been presented: the vector space approach and a cluster based approach using a Chinese Restaurant Process. Neither of the non-symbolic approaches to topic detection were able to consistently recognise the topics

Table 2: The distribution of topics across clusters following CRP training

Topic	Keyword	Clusters (no_in_cluster[cluser_id])
0	accommodation	77 [0] 1 [1] 59 [4] 13 [13] 1 [14] 1 [35] 6 [41] 37 [45] 1 [61] 2 [64] 2 [88]
1	bonus	122 [0] 1 [1] 14 [4] 11 [9] 2 [11] 9 [14] 27 [23] 2 [26] 2 [28] 10 [35]
2	break	22 [0] 161 [4] 2 [7] 2 [10] 2 [14] 1 [45] 1 [49] 1 [67] 5 [78] 1 [90] 1 [91] 1 [95]
3	deadline	142 [0] 33 [4] 2 [23] 5 [40] 7 [47] 3 [52] 2 [58] 3 [66] 3 [84]
4	deliverable	8 [0] 2 [4] 1 [14] 1 [45] 11 [86]
5	demotion	137 [0] 54 [4] 3 [14] 2 [23] 1 [29] 2 [39] 1 [45]
6	holiday	90 [0] 96 [4] 1 [6] 1 [14] 3 [24] 4 [25] 3 [42] 1 [61] 1 [78]
7	interview	123 [0] 75 [4] 1 [14] 1 [43]
8	management	134 [0] 12 [4] 2 [11] 3 [14] 4 [28] 1 [29] 3 [31] 4 [41] 26 [45] 1 [62] 2 [75] 4 [76] 1 [79] 3 [82]
9	meeting	148 [0] 38 [4] 1 [14] 5 [23] 1 [37] 1 [45] 3 [51] 1 [55] 1 [65] 1 [71]
10	move	78 [0] 51 [4] 56 [14] 1 [19] 1 [36] 1 [41] 1 [44] 1 [45] 1 [49] 2 [53] 2 [54] 1 [77] 3 [78] 1 [79]
11	occasion	91 [0] 83 [4] 1 [14] 2 [20] 1 [36] 2 [43] 6 [51] 5 [60] 2 [78] 7 [81]
12	perk	50 [0] 19 [1] 94 [4] 2 [7] 18 [14] 1 [15] 5 [41] 6 [45] 1 [68] 1 [71] 2 [93] 1 [94]
13	politics	128 [0] 35 [4] 7 [12] 1 [15] 2 [27] 3 [37] 21 [62] 2 [89] 1 [94]
14	presentation	132 [0] 32 [4] 2 [8] 2 [13] 17 [14] 1 [28] 2 [45] 1 [55] 1 [57] 1 [62] 1 [63] 2 [68] 2 [80] 4 [85]
15	project	121 [0] 43 [4] 1 [21] 4 [23] 1 [26] 1 [34] 2 [70] 1 [74] 23 [78] 1 [81] 2 [96]
16	promotion	71 [0] 1 [1] 55 [4] 58 [14] 1 [23] 1 [41] 2 [44] 1 [50] 1 [54] 1 [55] 3 [59] 2 [72] 1 [77] 2 [78]
17	report	147 [0] 33 [4] 2 [17] 1 [18] 1 [21] 3 [30] 4 [32] 4 [33] 1 [34] 1 [37] 2 [69] 1 [99]
18	review	134 [0] 1 [1] 1 [3] 30 [4] 13 [14] 4 [16] 1 [19] 11 [28] 1 [37] 2 [48] 1 [50] 1 [62]
19	technical	117 [0] 1 [2] 1 [3] 18 [4] 7 [14] 1 [38] 24 [45] 3 [56] 1 [57] 5 [62] 18 [68] 1 [74] 2 [92] 1 [97]
20	training	28 [0] 7 [1] 96 [4] 2 [5] 1 [6] 1 [8] 21 [14] 3 [22] 1 [28] 1 [33] 2 [38] 1 [41] 26 [45] 1 [49] 2 [65] 1 [71] 2 [73] 2 [83] 1 [92] 1 [98]
21	travel	90 [0] 59 [4] 2 [18] 1 [33] 1 [49] 1 [59] 1 [63] 1 [73] 38 [78] 6 [87]
22	weather	20 [0] 95 [4] 1 [7] 70 [14] 1 [15] 1 [59] 1 [67] 1 [72] 8 [78] 1 [90] 1 [91]
23	workload	72 [0] 1 [2] 114 [4] 8 [14] 2 [46] 1 [61] 1 [76] 1 [88]

Table 3: Clusters with 60% or more members from the same topic.

Cluster	Topics in cluster (keyword = no_extracts)
1	perk = 19 training = 7 review = 1 promotion = 1 bonus = 1 accommodation = 1
5	training = 2
9	bonus = 11
10	break = 2
12	politics = 7
13	accommodation = 13 presentation = 2
16	review = 4
17	report = 2
18	travel = 2 report = 1
20	occasion = 2
22	training = 3
23	bonus = 27 meeting = 5 project = 4 demotion = 2 deadline = 2 promotion = 1
24	holiday = 3
25	holiday = 4
27	politics = 2
30	report = 3
31	management = 3
32	report = 4
33	report = 4 travel = 1 training = 1
35	bonus = 10 accommodation = 1
39	demotion = 2
40	deadline = 5
42	holiday = 3
46	workload = 2
47	deadline = 7
48	review = 2
51	occasion = 6 meeting = 3
52	deadline = 3
53	move = 2
56	technical = 3
58	deadline = 2
59	promotion = 3 weather = 1 travel = 1
60	occasion = 5
62	politics = 21 technical = 5 review = 1 presentation = 1 management = 1
64	accommodation = 2
66	deadline = 3
68	technical = 18 presentation = 2 perk = 1
69	report = 2

Table 4: Clusters with 60% or more members from the same topic.

70	project = 2
75	management = 2
76	management = 4 workload = 1
80	presentation = 2
81	occasion = 7 project = 1
82	management = 3
83	training = 2
84	deadline = 3
85	presentation = 4
86	deliverable = 11
87	travel = 6
89	politics = 2
93	perk = 2
95	break = 1
96	project = 2
97	technical = 1
98	training = 1
99	report = 1

of user utterances. Furthermore, the non-symbolic approaches are unable to extract the relationships between events and their arguments and modifiers, as the NLU can. However, the non-symbolic approaches could provide a potential ‘repair’ mechanism to the Dialogue Manager whenever the NLU fails to produce an analysis of the user’s utterance, as can be the case when the utterances is particularly ungrammatical or if errors have been introduced through the ASR.

7 Appendix

```
Cluster 0:  77 [to0 S 0 0]  122 [to1 S 1 0]  22 [to2 S 2 0]  142 [to3 S 3 0]  8 [to4 S 4 0]
137 [to5 S 5 0]  90 [to6 S 6 0]  123 [to7 S 7 0]  134 [to8 S 8 0]  148 [to9 S 9 0]
78 [to10 S 10 0] 91 [to11 S 11 0] 50 [to12 S 12 0] 128 [to13 S 13 0] 132 [to14 S 14 0]
121 [to15 S 15 0] 71 [to16 S 16 0] 147 [to17 S 17 0] 134 [to18 S 18 0] 117 [to19 S 19 0]
28 [to20 S 20 0] 90 [to21 S 21 0] 20 [to22 S 22 0] 72 [to23 S 23 0]

Speaker/Level 1 pair counts:
S|meeting = 148 S|report = 147 S|deadline = 142 S|demotion = 137 S|review = 134
S|management = 134 S|presentation = 132 S|politics = 128 S|interview = 123 S|bonus = 122
S|project = 121 S|technical = 117 S|occasion = 91 S|travel = 90 S|holiday = 90
S|move = 78 S|accommodation = 77 S|workload = 72 S|promotion = 71 S|perk = 50
S|training = 28 S|break = 22 S|weather = 20 S|deliverable = 8

Cluster 1:  1 [to0 S 0 0]  1 [to1 S 1 0]  19 [to12 S 12 0] 1 [to16 S 16 0] 1 [to18 S 18 0]
7 [to20 S 20 0]

Speaker/Level 1 pair counts:
S|perk = 19 S|training = 7 S|review = 1 S|promotion = 1 S|bonus = 1
S|accommodation = 1

Cluster 2:  1 [to19 S 19 0] 1 [to23 S 23 0]

Speaker/Level 1 pair counts:
S|workload = 1 S|technical = 1

Cluster 3:  1 [to18 S 18 0] 1 [to19 S 19 0]

Speaker/Level 1 pair counts:
S|technical = 1 S|review = 1

Cluster 4:  59 [to0 S 0 0]  14 [to1 S 1 0]  161 [to2 S 2 0]  33 [to3 S 3 0]  2 [to4 S 4 0]
54 [to5 S 5 0]  96 [to6 S 6 0]  75 [to7 S 7 0]  12 [to8 S 8 0]  38 [to9 S 9 0]
51 [to10 S 10 0] 83 [to11 S 11 0] 94 [to12 S 12 0] 35 [to13 S 13 0] 32 [to14 S 14 0]
43 [to15 S 15 0] 55 [to16 S 16 0] 33 [to17 S 17 0] 30 [to18 S 18 0] 18 [to19 S 19 0]
96 [to20 S 20 0] 59 [to21 S 21 0] 95 [to22 S 22 0] 114 [to23 S 23 0]

Speaker/Level 1 pair counts:
S|break = 161 S|workload = 114 S|training = 96 S|holiday = 96 S|weather = 95
S|perk = 94 S|occasion = 83 S|interview = 75 S|travel = 59 S|accommodation = 59
S|promotion = 55 S|demotion = 54 S|move = 51 S|project = 43 S|meeting = 38
S|politics = 35 S|report = 33 S|deadline = 33 S|presentation = 32 S|review = 30
S|technical = 18 S|bonus = 14 S|management = 12 S|deliverable = 2

Cluster 5:  2 [to20 S 20 0]

Speaker/Level 1 pair counts:
S|training = 2

Cluster 6:  1 [to6 S 6 0]  1 [to20 S 20 0]

Speaker/Level 1 pair counts:
S|training = 1 S|holiday = 1

Cluster 7:  2 [to2 S 2 0]  2 [to12 S 12 0] 1 [to22 S 22 0]

Speaker/Level 1 pair counts:
S|perk = 2 S|break = 2 S|weather = 1

Cluster 8:  2 [to14 S 14 0] 1 [to20 S 20 0]

Speaker/Level 1 pair counts:
S|presentation = 2 S|training = 1
```

```
Cluster 9:  11 [to1 S 1 0]

Speaker/Level 1 pair counts:
S|bonus = 11

Cluster 10:  2 [to2 S 2 0]

Speaker/Level 1 pair counts:
S|break = 2

Cluster 11:  2 [to1 S 1 0]  2 [to8 S 8 0]

Speaker/Level 1 pair counts:
S|management = 2 S|bonus = 2

Cluster 12:  7 [to13 S 13 0]

Speaker/Level 1 pair counts:
S|politics = 7

Cluster 13:  13 [to0 S 0 0]  2 [to14 S 14 0]

Speaker/Level 1 pair counts:
S|accommodation = 13 S|presentation = 2

Cluster 14:  1 [to0 S 0 0]  9 [to1 S 1 0]  2 [to2 S 2 0]  1 [to4 S 4 0]  3 [to5 S 5 0]
1 [to6 S 6 0]  1 [to7 S 7 0]  3 [to8 S 8 0]  1 [to9 S 9 0]  56 [to10 S 10 0]
1 [to11 S 11 0] 18 [to12 S 12 0] 17 [to14 S 14 0] 58 [to16 S 16 0] 13 [to18 S 18 0]
7 [to19 S 19 0] 21 [to20 S 20 0] 70 [to22 S 22 0] 8 [to23 S 23 0]

Speaker/Level 1 pair counts:
S|weather = 70 S|promotion = 58 S|move = 56 S|training = 21 S|perk = 18
S|presentation = 17 S|review = 13 S|bonus = 9 S|workload = 8 S|technical = 7
S|management = 3 S|demotion = 3 S|break = 2 S|occasion = 1 S|meeting = 1
S|interview = 1 S|holiday = 1 S|deliverable = 1 S|accommodation = 1

Cluster 15:  1 [to12 S 12 0] 1 [to13 S 13 0] 1 [to22 S 22 0]

Speaker/Level 1 pair counts:
S|weather = 1 S|politics = 1 S|perk = 1

Cluster 16:  4 [to18 S 18 0]

Speaker/Level 1 pair counts:
S|review = 4

Cluster 17:  2 [to17 S 17 0]

Speaker/Level 1 pair counts:
S|report = 2

Cluster 18:  1 [to17 S 17 0] 2 [to21 S 21 0]

Speaker/Level 1 pair counts:
S|travel = 2 S|report = 1

Cluster 19:  1 [to10 S 10 0] 1 [to18 S 18 0]

Speaker/Level 1 pair counts:
S|review = 1 S|move = 1

Cluster 20:  2 [to11 S 11 0]

Speaker/Level 1 pair counts:
```

```
S|occasion = 2

Cluster 21:  1 [to15 S 15 0] 1 [to17 S 17 0]

Speaker/Level 1 pair counts:
S|report = 1 S|project = 1

Cluster 22:  3 [to20 S 20 0]

Speaker/Level 1 pair counts:
S|training = 3

Cluster 23:  27 [to1 S 1 0]  2 [to3 S 3 0]  2 [to5 S 5 0]  5 [to9 S 9 0]  4 [to15 S 15 0]
1 [to16 S 16 0]

Speaker/Level 1 pair counts:
S|bonus = 27 S|meeting = 5 S|project = 4 S|demotion = 2 S|deadline = 2
S|promotion = 1

Cluster 24:  3 [to6 S 6 0]

Speaker/Level 1 pair counts:
S|holiday = 3

Cluster 25:  4 [to6 S 6 0]

Speaker/Level 1 pair counts:
S|holiday = 4

Cluster 26:  2 [to1 S 1 0]  1 [to15 S 15 0]

Speaker/Level 1 pair counts:
S|bonus = 2 S|project = 1

Cluster 27:  2 [to13 S 13 0]

Speaker/Level 1 pair counts:
S|politics = 2

Cluster 28:  2 [to1 S 1 0]  4 [to8 S 8 0]  1 [to14 S 14 0] 11 [to18 S 18 0] 1 [to20 S 20 0]


Speaker/Level 1 pair counts:
S|review = 11 S|management = 4 S|bonus = 2 S|training = 1 S|presentation = 1


Cluster 29:  1 [to5 S 5 0]  1 [to8 S 8 0]

Speaker/Level 1 pair counts:
S|management = 1 S|demotion = 1

Cluster 30:  3 [to17 S 17 0]

Speaker/Level 1 pair counts:
S|report = 3

Cluster 31:  3 [to8 S 8 0]

Speaker/Level 1 pair counts:
S|management = 3

Cluster 32:  4 [to17 S 17 0]

Speaker/Level 1 pair counts:
S|report = 4
```

```
Cluster 33:  4 [to17 S 17 0] 1 [to20 S 20 0] 1 [to21 S 21 0]

Speaker/Level 1 pair counts:
S|report = 4 S|travel = 1 S|training = 1

Cluster 34:  1 [to15 S 15 0] 1 [to17 S 17 0]

Speaker/Level 1 pair counts:
S|report = 1 S|project = 1

Cluster 35:  1 [to0 S 0 0]  10 [to1 S 1 0]

Speaker/Level 1 pair counts:
S|bonus = 10 S|accommodation = 1

Cluster 36:  1 [to10 S 10 0] 1 [to11 S 11 0]

Speaker/Level 1 pair counts:
S|occasion = 1 S|move = 1

Cluster 37:  1 [to9 S 9 0]  3 [to13 S 13 0] 1 [to17 S 17 0] 1 [to18 S 18 0]

Speaker/Level 1 pair counts:
S|politics = 3 S|review = 1 S|report = 1 S|meeting = 1

Cluster 38:  1 [to19 S 19 0] 2 [to20 S 20 0]

Speaker/Level 1 pair counts:
S|training = 2 S|technical = 1

Cluster 39:  2 [to5 S 5 0]

Speaker/Level 1 pair counts:
S|demotion = 2

Cluster 40:  5 [to3 S 3 0]

Speaker/Level 1 pair counts:
S|deadline = 5

Cluster 41:  6 [to0 S 0 0]  4 [to8 S 8 0]  1 [to10 S 10 0] 5 [to12 S 12 0] 1 [to16 S 16 0]
1 [to20 S 20 0]

Speaker/Level 1 pair counts:
S|accommodation = 6 S|perk = 5 S|management = 4 S|training = 1 S|promotion = 1
S|move = 1

Cluster 42:  3 [to6 S 6 0]

Speaker/Level 1 pair counts:
S|holiday = 3

Cluster 43:  1 [to7 S 7 0]  2 [to11 S 11 0]

Speaker/Level 1 pair counts:
S|occasion = 2 S|interview = 1

Cluster 44:  1 [to10 S 10 0] 2 [to16 S 16 0]

Speaker/Level 1 pair counts:
S|promotion = 2 S|move = 1

Cluster 45:  37 [to0 S 0 0]  1 [to2 S 2 0]  1 [to4 S 4 0]  1 [to5 S 5 0]  26 [to8 S 8 0]
1 [to9 S 9 0]  1 [to10 S 10 0] 6 [to12 S 12 0] 2 [to14 S 14 0] 24 [to19 S 19 0]
```

```
26 [to20 S 20 0]

Speaker/Level 1 pair counts:
S|accommodation = 37 S|training = 26 S|management = 26 S|technical = 24 S|perk = 6
S|presentation = 2 S|move = 1 S|meeting = 1 S|demotion = 1 S|deliverable = 1
S|break = 1

Cluster 46:  2 [to23 S 23 0]

Speaker/Level 1 pair counts:
S|workload = 2

Cluster 47:  7 [to3 S 3 0]

Speaker/Level 1 pair counts:
S|deadline = 7

Cluster 48:  2 [to18 S 18 0]

Speaker/Level 1 pair counts:
S|review = 2

Cluster 49:  1 [to2 S 2 0]  1 [to10 S 10 0] 1 [to20 S 20 0] 1 [to21 S 21 0]

Speaker/Level 1 pair counts:
S|travel = 1 S|training = 1 S|move = 1 S|break = 1

Cluster 50:  1 [to16 S 16 0] 1 [to18 S 18 0]

Speaker/Level 1 pair counts:
S|review = 1 S|promotion = 1

Cluster 51:  3 [to9 S 9 0]  6 [to11 S 11 0]

Speaker/Level 1 pair counts:
S|occasion = 6 S|meeting = 3

Cluster 52:  3 [to3 S 3 0]

Speaker/Level 1 pair counts:
S|deadline = 3

Cluster 53:  2 [to10 S 10 0]

Speaker/Level 1 pair counts:
S|move = 2

Cluster 54:  2 [to10 S 10 0] 1 [to16 S 16 0]

Speaker/Level 1 pair counts:
S|move = 2 S|promotion = 1

Cluster 55:  1 [to9 S 9 0]  1 [to14 S 14 0] 1 [to16 S 16 0]

Speaker/Level 1 pair counts:
S|promotion = 1 S|presentation = 1 S|meeting = 1

Cluster 56:  3 [to19 S 19 0]

Speaker/Level 1 pair counts:
S|technical = 3

Cluster 57:  1 [to14 S 14 0] 1 [to19 S 19 0]

Speaker/Level 1 pair counts:
```

```
S|technical = 1 S|presentation = 1

Cluster 58:  2 [to3 S 3 0]

Speaker/Level 1 pair counts:
S|deadline = 2

Cluster 59:  3 [to16 S 16 0] 1 [to21 S 21 0] 1 [to22 S 22 0]

Speaker/Level 1 pair counts:
S|promotion = 3 S|weather = 1 S|travel = 1

Cluster 60:  5 [to11 S 11 0]

Speaker/Level 1 pair counts:
S|occasion = 5

Cluster 61:  1 [to0 S 0 0]  1 [to6 S 6 0]  1 [to23 S 23 0]

Speaker/Level 1 pair counts:
S|workload = 1 S|holiday = 1 S|accommodation = 1

Cluster 62:  1 [to8 S 8 0]  21 [to13 S 13 0] 1 [to14 S 14 0] 1 [to18 S 18 0] 5 [to19 S 19 0]


Speaker/Level 1 pair counts:
S|politics = 21 S|technical = 5 S|review = 1 S|presentation = 1 S|management = 1


Cluster 63:  1 [to14 S 14 0] 1 [to21 S 21 0]

Speaker/Level 1 pair counts:
S|travel = 1 S|presentation = 1

Cluster 64:  2 [to0 S 0 0]

Speaker/Level 1 pair counts:
S|accommodation = 2

Cluster 65:  1 [to9 S 9 0]  2 [to20 S 20 0]

Speaker/Level 1 pair counts:
S|training = 2 S|meeting = 1

Cluster 66:  3 [to3 S 3 0]

Speaker/Level 1 pair counts:
S|deadline = 3

Cluster 67:  1 [to2 S 2 0]  1 [to22 S 22 0]

Speaker/Level 1 pair counts:
S|weather = 1 S|break = 1

Cluster 68:  1 [to12 S 12 0] 2 [to14 S 14 0] 18 [to19 S 19 0]

Speaker/Level 1 pair counts:
S|technical = 18 S|presentation = 2 S|perk = 1

Cluster 69:  2 [to17 S 17 0]

Speaker/Level 1 pair counts:
S|report = 2

Cluster 70:  2 [to15 S 15 0]
```

```
Speaker/Level 1 pair counts:
S|project = 2

Cluster 71:  1 [to9 S 9 0]  1 [to12 S 12 0] 1 [to20 S 20 0]

Speaker/Level 1 pair counts:
S|training = 1 S|perk = 1 S|meeting = 1

Cluster 72:  2 [to16 S 16 0] 1 [to22 S 22 0]

Speaker/Level 1 pair counts:
S|promotion = 2 S|weather = 1

Cluster 73:  2 [to20 S 20 0] 1 [to21 S 21 0]

Speaker/Level 1 pair counts:
S|training = 2 S|travel = 1

Cluster 74:  1 [to15 S 15 0] 1 [to19 S 19 0]

Speaker/Level 1 pair counts:
S|technical = 1 S|project = 1

Cluster 75:  2 [to8 S 8 0]

Speaker/Level 1 pair counts:
S|management = 2

Cluster 76:  4 [to8 S 8 0]  1 [to23 S 23 0]

Speaker/Level 1 pair counts:
S|management = 4 S|workload = 1

Cluster 77:  1 [to10 S 10 0] 1 [to16 S 16 0]

Speaker/Level 1 pair counts:
S|promotion = 1 S|move = 1

Cluster 78:  5 [to2 S 2 0]  1 [to6 S 6 0]  3 [to10 S 10 0] 2 [to11 S 11 0] 23 [to15 S 15 0]
2 [to16 S 16 0] 38 [to21 S 21 0] 8 [to22 S 22 0]

Speaker/Level 1 pair counts:
S|travel = 38 S|project = 23 S|weather = 8 S|break = 5 S|move = 3
S|promotion = 2 S|occasion = 2 S|holiday = 1

Cluster 79:  1 [to8 S 8 0]  1 [to10 S 10 0]

Speaker/Level 1 pair counts:
S|move = 1 S|management = 1

Cluster 80:  2 [to14 S 14 0]

Speaker/Level 1 pair counts:
S|presentation = 2

Cluster 81:  7 [to11 S 11 0] 1 [to15 S 15 0]

Speaker/Level 1 pair counts:
S|occasion = 7 S|project = 1

Cluster 82:  3 [to8 S 8 0]

Speaker/Level 1 pair counts:
S|management = 3
```

```
Cluster 83:  2 [to20 S 20 0]

Speaker/Level 1 pair counts:
S|training = 2

Cluster 84:  3 [to3 S 3 0]

Speaker/Level 1 pair counts:
S|deadline = 3

Cluster 85:  4 [to14 S 14 0]

Speaker/Level 1 pair counts:
S|presentation = 4

Cluster 86:  11 [to4 S 4 0]

Speaker/Level 1 pair counts:
S|deliverable = 11

Cluster 87:  6 [to21 S 21 0]

Speaker/Level 1 pair counts:
S|travel = 6

Cluster 88:  2 [to0 S 0 0]  1 [to23 S 23 0]

Speaker/Level 1 pair counts:
S|accommodation = 2 S|workload = 1

Cluster 89:  2 [to13 S 13 0]

Speaker/Level 1 pair counts:
S|politics = 2

Cluster 90:  1 [to2 S 2 0]  1 [to22 S 22 0]

Speaker/Level 1 pair counts:
S|weather = 1 S|break = 1

Cluster 91:  1 [to2 S 2 0]  1 [to22 S 22 0]

Speaker/Level 1 pair counts:
S|weather = 1 S|break = 1

Cluster 92:  2 [to19 S 19 0] 1 [to20 S 20 0]

Speaker/Level 1 pair counts:
S|technical = 2 S|training = 1

Cluster 93:  2 [to12 S 12 0]

Speaker/Level 1 pair counts:
S|perk = 2

Cluster 94:  1 [to12 S 12 0] 1 [to13 S 13 0]

Speaker/Level 1 pair counts:
S|politics = 1 S|perk = 1

Cluster 95:  1 [to2 S 2 0]

Speaker/Level 1 pair counts:
S|break = 1
```

```
Cluster 96:  2 [to15 S 15 0]

Speaker/Level 1 pair counts:
S|project = 2

Cluster 97:  1 [to19 S 19 0]

Speaker/Level 1 pair counts:
S|technical = 1

Cluster 98:  1 [to20 S 20 0]

Speaker/Level 1 pair counts:
S|training = 1

Cluster 99:  1 [to17 S 17 0]

Speaker/Level 1 pair counts:
S|report = 1
```

References

[1] C.E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[2] Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier. *Modern Information Retrieval*. Addison Wesley, 1st edition, May 1999.

[3] Becker, J. and Kuropka, D. Topic-based Vector Space Model. In *Proceedings of the 6th International Conference on Business Information Systems*, pages 7–12, Colorado Springs, July 2003.

[4] M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

[5] R. Granell, S. Pulman, C.D. Martínez, and J.M. Benedí. Dialogue act tagging and segmentation with a single perceptron. In *Interspeech: 11th Annual Conference of the International Speech Communication Association*, pages 3074–3077, Makuhari, Japan, 2010.

[6] Arne Jönsson, Frida Andén, Lars Degerstedt, Annika Flycht-Eriksson, Magnus Merkel, and Sara Norberg. Experiences from combining dialogue system development with information extraction techniques. In Mark T. Maybury, editor, *New Directions in Question Answering*. AAAI Press/MIT Press, 2004.

[7] Christopher Kennedy and Branimir Boguraev. Anaphora for everyone: pronominal anaphora resoluation without a parser. In *Proceedings of the 16th conference on Computational linguistics*, pages 113–118, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

[8] S.N. Maceachern and P. Müller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

[9] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008.

[10] Antonio Molina and Ferran Pla. Shallow parsing using specialized HMMs. *Journal of Machine Learning Research*, 2:595–613, 2002.

[11] J. Pitman. Combinatorial stochastic processes, 2002.

[12] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997.

[13] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[14] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, March 2002.