Combining Hidden Markov Models and Latent Semantic Analysis for Topic Segmentation and Labeling: Method and Clinical Application

Filip Ginter,[1] **Hanna Suominen,**[1,2] **Sampo Pyysalo**[2] and **Tapio Salakoski**[1,2]
[1]Department of Information Technology, University of Turku and
[2]Turku Centre for Computer Science (TUCS)
Joukahaisenkatu 3-5,
20520 Turku, Finland
first.last@utu.fi

Abstract

Topic segmentation and labeling systems enable fine-grained information search. However, previously proposed methods require annotated data to adapt to different information needs and have limited applicability to texts with short segment length. We introduce an unsupervised method based on a combination of Hidden Markov Models and latent semantic indexing which allows the topics of interest to be defined freely, without the need for data annotation, and can identify short segments. The method is evaluated in an application domain of intensive care nursing narratives. It is shown to considerably outperform a keyword-based heuristic baseline and to achieve a level of performance comparable to that of a related supervised method trained on 3600 manually annotated words.

1 Introduction

We have previously introduced an application of Hidden Markov Models (HMMs) to topic segmentation (TS) and labeling of Finnish intensive care unit (ICU) nursing narratives (Suominen et al., 2008). In this application, common and repeatedly discussed topics, such as breathing and hemodynamics, are identified in the text, supporting information access and clinical decision-making. In this supervised approach, annotated training data are necessary to induce the HMM model and consequently, the set of possible topics cannot be changed without annotation of additional training data.

In this paper, we introduce a topic segmentation and labeling method where the set of possible topics is not predetermined but is provided by the user as a set of freely chosen keywords, such as *breathing* or *hemodynamics*. The proposed method does not require labeled training data and is, in this respect, unsupervised. This property allows the topics of interest to be easily changed — the user simply specifies new keywords — whereas for a supervised TS and labeling system a new training set would need to be annotated.

The proposed method is a combination of latent semantic analysis (LSA) and a graphical model closely related to HMMs. The method is particularly suitable in cases where almost all documents contain relevant information about the given topics, and the topic segments are short, even shorter than a single sentence. The applicability of existing unsupervised TS methods in these cases is likely to be limited. On the other hand, supervised methods relying on manually labeled training data cannot be applied when the topics can be chosen freely.

Our motivation and scope comes here from ICU narratives. However, we believe that as a general TS and labeling technique supporting *ad hoc* information needs the introduced method may find application also in many other, unrelated domains. As an example of a class of texts which are also characterized by short, unmarked segments, consider scientific publication abstracts, where the method could be applied e.g. to separate between *methods* and *results*-related segments.

2 Related work

TS (alternatively referred to as text segmentation), the automatic division of text into topically coherent units, is a well-studied problem. Many

TS methods are based on the location of first uses of word types, pronoun reference, punctuation marks or other linguistic cues implying topic-change boundaries. The cues are either hard-coded domain-specific rules or induced by machine learning from a corpus (Beeferman et al., 1997; Reynar, 1999).

Another common approach to TS is to consider the similarity of text before and after a proposed segment boundary by measuring, for example, word co-occurrence, repetition or semantic relations; a sudden drop in similarity indicates a likely change in topic. Algorithms based on this approach can be fully unsupervised (Hearst, 1997; Ferret, 2002). Further, LSA has been shown to improve TS when used as a text similarity measure (Bestgen, 2006).

A third major group of TS methods is based on graphical models for sequence labeling. For instance, HMMs have been applied (Yamron et al., 1998; Blei and Moreno, 2001; Suominen et al., 2008). These methods are supervised, but otherwise resemble ours; the approach is a natural choice because segmentation is given by the assigned topic labels.

The applicability of existing TS systems is, however, limited in our case. To allow a free choice of topics of interest, we aim at an unsupervised approach. Further, our data is characterized by very short segment length — several topic changes may occur within a single sentence. Existing unsupervised TS methods require considerably longer segment sizes (see, e.g., (Hearst, 1997; Ferret, 2002)) to reliably detect topic change boundaries. For instance, the TextTiling method of Hearst (1997) searches for topic boundaries between contexts of 200 tokens, whereas the average topic length in our data was only 18 tokens, that is, an order of magnitude shorter. For short texts, techniques similar to query expansion in information extraction and use of likely topic length have been proposed (Ponte and Croft, 1997; Chang and Lee, 2003), but these studies do not, however, consider topic labeling.

In our application domain, Cho et al. (2003) have applied TS and labeling to medical narratives from radiology and urology departments. However, their method relies strongly on hard-coded headlining rules, linguistic cues and lexical patterns seen within training examples. TS techniques have also been designed for the temporal order analysis of medical discharge summaries using a statistical parser to segment the sentences into clauses and two supervised classifiers to predict the segment boundaries and assign for every segment pair their time-wise order (Bramsen et al., 2006). Finally, Hiissa et al. (2007) have introduced a supervised system classifying segments of intensive care patient narratives with respect to topics of *breathing*, *blood circulation*, and *pain*; the segments were, however, created manually.

3 Patient documentation data

The data used in this study consists of nursing notes of 516 adult ICU patients. These Finnish patient-specific records are written during every shift and are mainly used for intra-unit information exchange.

The data set consists of 17140 nursing shifts. We apply a simple domain-adapted tokenizer, obtaining 1.2 million tokens (including punctuation). Each shift thus contains, on average, 73 tokens. The most common topics of the text were *breathing*, *hemodynamics*, *consciousness*, *relatives*, and *diuresis*. Approximately half of the shifts contain explicit topic headings, although these are not standardized and are often misspelled or abbreviated. Additionally, the text is often telegraphic and the vocabulary is highly specialized with a substantial amount of professional terminology, unit-specific documentation practices, and frequent misspellings. Figure 1 illustrates the data.

As test data, we use a manually annotated subset consisting of 402 shifts randomly chosen from the records of first 135 patients by their admission date (Suominen et al., 2008). In the annotation we identify segments belonging to the topics listed above; text not belonging to any of these is assigned the topic *other*. The average length of a topic segment is 18 tokens.

4 Method

We now first recall basic notions of LSA and HMMs and then proceed to introduce the unsupervised TS and labeling method which is based on their combination. The main insight of the proposed method is that the LSA similarity of words to the given topic keywords can be used to replace HMM emission probabilities. Whereas a supervised HMM requires labeled data to estimate the

a)
Night shift
B R E A T H I N G: Doing nicely with the mask. Smallish carbondioxide retention after pain killers, otherwise CO2 < 8.Hourly breathing exercises. Mucus -> wheezing. Able to cough faintly&swallow mucus.
C O N S C I O U S N E S S: Spontaneously awake. DRUGNAME 5mg i.m. After that, was able to nod peacefully. Copes the breathing exercices so-and-so. The strenght in the extremities exept the left hand with bandage are weak.
H e m o d : RRmap staying >65. In a sleep quite low RR.Reduced amount of DRUGNAME and full stop in the small huors. Steady SR.
D I U R E S I S : More DRUGNAME -> Diuresis > 150 ml/h. Fuzzy in the evening.
O T H E R : Son and his wife visiting.Hevay moistening to mouth.
2006-02-01 04:55

b)
Long morning s
After admission fast FA which we treid to invert with electrcity (x3) without result. later FA freq extremely varying and quite economic. After 14 o'clock, pulse occasionally tachycardic, slowed down with DRUGNAME and DRUGNAME infusion (load 150 mg, maintenance 1200 mg/day). Inversion to SR at about 17.30.Hemodynamics quite stable, DRUGNAMEinfusion cont with moder dosage.
Diuresis narrow, morning DRUGNAME.
PCWP highish (21). Adequate CI.
Dr flow normal, narrow.
Forenoon: despite medicatio, tried to breath 'against respirator', which is the reason for relaxation (a couple of times).
Own breatthing started and woke up regardless of sedation & kooperative. With CPAP ok ox and ventilation.
2006-12-11 18:02

Figure 1: Example of Finnish nursing notes translated to English preserving all typing errors and typographical properties. The Finnish originals are not included due to space considerations. Note the topic headings in report *a* with the untypical use of the heading *other* instead of the more common *relatives*. In contrast to *a*, the report *b* does not contain explicit topic headings.

emission probabilities, the unsupervised method only requires a single keyword for each topic.

4.1 Latent semantic analysis

LSA is a commonly applied technique for inducing text similarity measures from co-occurrence statistics in a large, unannotated corpus of text. In our case, we use an LSA-based term-term similarity measure. The standard LSA method based on decomposition of the term-by-document matrix is not applicable because the context in which it measures word co-occurrence is the whole document. In our case, however, the topic keywords occur in the majority of documents — here document refers to a single shift — and, more importantly, different topics tend to co-occur in a single document, therefore not allowing document-level distribution of terms to sufficiently distinguish the various topics. Instead, we apply the Word Space model (Schütze, 1998) which decomposes a term-by-term matrix and only considers word co-occurrence within a fixed context window rather than in the whole document, therefore allowing sub-document distributional properties to be accounted for.

We denote the LSA similarity of word $w_j, j \in \{1, \ldots N_w\}$, to topic $q_i, i \in \{1, \ldots N_q\}$, as $lsa(w_j, q_i)$. Here N_w is the vocabulary size, N_q is the number of possible topics, and q_i is the keyword specified by the user for the respective topic. In our experiments, we use the Finnish equivalents of the keywords *breathing*, *hemodynamics*, *consciousness*, *relative* and *diuresis* to define the five annotated topics. The sixth topic, *other*, is characterized as an LSA query *other NOT breathing NOT hemodynamics NOT consciousness NOT relative NOT diuresis*. The negation operator *NOT* is available in Word Space LSA queries (Widdows and Peters, 2003). The resulting LSA scores are illustrated in Figure 2; they are obtained by first performing LSA on unannotated ICU narrative texts and then calculating the LSA similarity of each vocabulary word with the respective topic keyword (or LSA query with negations). Punctuation, numbers, and small number of extremely common stop-words are excluded from the LSA calculation.

4.2 Hidden Markov Models

We model the problem of segmenting the clinical texts and assigning a topic to each resulting segment as a sequence labeling task. Given an input word sequence $w = (w(1), \ldots, w(T))$, each word $w(t)$, $t \in \{1, \ldots, T\}$, is assigned a topic label $q(t) \in \{q_1, \ldots, q_{N_q}\}$. Each word $w(t)$ belongs to the vocabulary $\{w_1, \ldots, w_{N_w}\}$.

The sequence labeling problem can be solved by an HMM with N_q states where w corresponds to the visible sequence of observations and the sequence of labels $q = (q(1), \ldots, q(T))$ corresponds to the hidden sequence of HMM states. We use a first-order HMM, thus a particular hidden variable $q(t)$ only depends on the previous hidden state $q(t-1)$, and an observed variable

RELATIVES		HEMODYNAMICS		OTHER	
relative	1.000	hemodynamics	1.000	stomach	0.683
phone	0.947	pulse	0.910	other	0.682
daughter	0.916	sr	0.819	net	0.676
wife	0.889	rr-level	0.785	hemolyzed	0.673
visit	0.877	highish	0.784	shirt	0.637
son	0.859	sinus_rythm	0.784	contrast_medium_boosted	0.635
watch	0.821	rr	0.768	blanket	0.630
husband	0.820	blood_pressure	0.716	from_DRUGNAME	0.618
brother	0.785	extrasystole	0.673	soft	0.618
sister	0.777	ok	0.672	puncture_sample	0.614

Figure 2: Translated examples of the words most similar to selected topics and their associated LSA similarity values.

$w(t)$ is only dependent on the value of the hidden variable $q(t)$. Additionally, the initial probability of states is uniformly distributed. The labeling given by the HMM is the best hidden state sequence $\hat{q}$ obtained by solving

$$\hat{q} = \arg\max_{q \in \mathcal{Q}} P(w, q), \tag{1}$$

where $\mathcal{Q}$ is the space of all hidden state sequences and

$$P(w, q) = P(w(1)|q(1)) \cdot \prod_{t=2}^{T} P(w(t)|q(t))P(q(t)|q(t-1)).$$

The optimal sequence $\hat{q}$ is known as the Viterbi path and the optimization problem (1) can be efficiently computed using the standard Viterbi algorithm. For a detailed introduction to these algorithms, see, for example, Rabiner (1989).

4.3 The proposed unsupervised method

In order to solve (1), the conditional probabilities $P(w(t)|q(t))$, typically referred to as the *emission probabilities*, and $P(q(t)|q(t-1))$, typically referred to as the *transition probabilities*, must be defined. In the supervised case, these are obtained from training data as maximum-likelihood estimates. Here we aim to obtain these conditional probabilities in a minimally-supervised manner which does not require annotated training data. To simplify the notation, we will refer in the following text, whenever possible, to the conditional probabilities $P(w_j|q_i)$ and $P(q_j|q_i)$ without the sequence index t.

4.3.1 Transition probabilities $P(q_j|q_i)$

We distribute the transition probabilities uniformly since, due to our unsupervised setting, there is no annotated data available for direct estimation. In order to be able to control the likelihood of switching from one topic to another, thus controlling the segmentation granularity, we introduce a *self-transition probability* parameter $\delta \in (0, 1)$. The HMM transition probability is then defined as

$$P(q_j|q_i) = \begin{cases} \delta & \text{if } j = i \\ \frac{1-\delta}{N_q - 1} & \text{if } j \neq i \end{cases}.$$

The probability of continuing the current topic is thus δ, and the remaining probability $1 - \delta$ of switching a topic is distributed evenly. Trivially, $\sum_{q_j} P(q_j|q_i) = 1$ for any q_i.

4.3.2 Emission probabilities $P(w_j|q_i)$

Our aim is to derive the value of the emission probability $P(w_j|q_i)$ from the LSA similarity $lsa(w_j, q_i)$ of the word w_j to the topic q_i, or more accurately to the keyword that defines the topic q_i. A straightforward approach is to normalize the LSA similarity into probabilities so that

$$P(w_j|q_i) = \frac{lsa(w_j, q_i)}{\sum_{k=1}^{N_w} lsa(w_k, q_i)}. \tag{2}$$

This normalization strategy, however, assumes that there is some total mass of relatedness to be redistributed by LSA among the individual words and that this mass is topic-independent. Otherwise, a topic with a small number of related terms will distribute the probability mass of 1 among a small number of words as opposed to a topic with a large number of related terms. Consequently, the emission probabilities of such a topic will numerically dominate the calculation of the Viterbi path $\hat{q}$ and result in poor performance of the model — an effect we have observed in our early experiments. We avoid this type of numerical domination by relaxing the HMM model.

4.3.3 Relaxed graphical model

Instead of normalizing the LSA similarities by Equation 2, we use the unnormalized LSA values directly. This yields a graphical model that preserves the overall structure of an HMM but replaces the emission probabilities with a quantity that is not a probability. The optimal state sequence in this graphical model is then obtained by solving $\arg\max_{q \in \mathcal{Q}} C(w, q)$, where

$$C(w,q) = lsa(w(1), q(1)) \cdot \prod_{t=2}^{T} lsa(w(t), q(t)) P(q(t)|q(t-1)).$$

Replacing the probability $P(w(t)|q(t))$ with the non-probability $lsa(w(t), q(t))$ is the only difference between the HMM cost function $P(w, q)$ and the relaxed model cost function $C(w, q)$. This change does not violate any assumptions in the Viterbi algorithm which thus remains directly applicable to the computation of the optimal sequence of states also in the relaxed model.

This relaxed formalization does not suffer from the problem of a single topic numerically dominating the cost function value and, in our preliminary experiments, resulted in a significant gain in performance. However, a problem of mutual incomparability of the LSA similarity values across topics persists; there is no basis for the implicit assumption that the same LSA similarity value corresponds to the same underlying degree of relatedness, regardless of the topic in question. As an illustrative example of the general problem, let us consider a topic q_1 defined by a single keyword u_1. We then have $lsa(u_1, q_1) = 1$ since the LSA similarity of a word to itself is by definition 1. On the other hand, this does not hold for topics defined by more than one keyword, where the similarity of any of the several defining keywords with the topic is strictly smaller than 1 (except in degenerate cases). Consequently, the same degree of relatedness does not necessarily correspond to the same LSA similarity values across topics. A re-scaling strategy is thus called for which would aim to improve the numerical comparability of the LSA values across topics. We introduce one such possible strategy based on the following insight.

Let us consider words in the descending order by their LSA similarity to a topic q_i and compare for each word its LSA similarity with q_i and the maximum of its LSA similarities with any topic other than q_i (see Figure 3 for illustration). The position in the ordering at which, for the first time, a word has a higher similarity with a topic other than q_i, which we refer to as the *impact index*, naturally divides the ordered list of words into two parts. Words up to the impact index are those that have a high LSA similarity to the topic q_i and, at the same time, do not have higher similarity with any other topic. These words are thus strong indicators of the topic q_i. The LSA similarity of the word at the impact index, which we refer to as the *impact similarity* is then, for the topic q_i, a natural cut-off point that gives the lowest LSA similarity at which the words can yet be considered as strong indicators of the topic. Numerically, the impact index and impact index similarity may vary significantly across topics.

Since the impact similarity has a clear intuitive interpretation, we propose a strategy which re-scales the LSA values for each topic so that the impact similarity is set to a given, topic-independent constant α. Additionally, the re-scaling sets the LSA similarity of the most similar word for any topic as equal to 1 and the minimal similarity of any word to any topic to be a constant β. The effect of this re-scaling is illustrated in Figure 3.

We now proceed to define the re-scaling strategy formally. Let us consider an ordering π_i of the words such that the value $\pi_i(w_j)$ gives the index at which the word w_j is found in a sequence of words ordered in descending order by their LSA similarity with q_i. Let $lsa_1(q_i) = \max_{w_j} lsa(w_j, q_i)$ and $lsa_m(q_i) = \min_{w_j} lsa(w_j, q_i)$. Finally, let $lsa_I(q_i)$ denote the LSA similarity $lsa(w_j, q_i)$ where $\pi_i(w_j) = I(q_i)$, that is, the impact point similarity for topic q_i. These quantities are illustrated in Figure 3. The re-scaled LSA similarity, denoted $\overline{lsa}$, is then defined in Equation 3.

The optimal state sequence through our final model is then obtained by solving $\arg\max_{q \in \mathcal{Q}} \overline{C}(w, q)$, where

$$\overline{C}(w,q) = \overline{lsa}(w(1), q(1)) \cdot \prod_{t=2}^{T} \overline{lsa}(w(t), q(t)) P(q(t)|q(t-1)).$$

To summarize, we have now obtained a graphical model for unsupervised topic segmentation and labeling of text that is closely related to first-order HMMs. The transition probabilities other

$$\overline{lsa}(w_j, q_i) = \begin{cases} \frac{1-\alpha}{lsa_1(q_i)-lsa_I(q_i)} \cdot (lsa(w_j, q_i) - lsa_I(q_i)) + \alpha & \text{if } \pi_i(w_j) \leq I(q_i) \\ \frac{\alpha-\beta}{lsa_I(q_i)-lsa_m(q_i)} \cdot (lsa(w_j, q_i) - lsa_m(q_i)) + \beta & \text{otherwise} \end{cases} \quad (3)$$

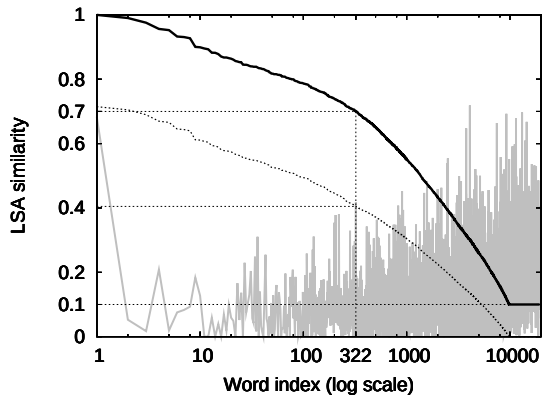

Figure 3: The effects of re-scaling the LSA similarity values of the topic *other* by Equation 3. The re-scaled LSA values are shown as a full line, the unscaled LSA values as a dotted line, and the maximum LSA similarity with any other topic as a gray line. The important characteristics of the LSA values in this case are: $lsa_1(q_i) = 0.71$, $I(q_i) = 322$, $lsa_I(q_i) = 0.41$, and $lsa_m(q_i) = 0$. The re-scaling parameters are $\alpha = 0.7$ and $\beta = 0.1$.

than the parameterized self-transition probability δ are uniformly distributed and the emission probabilities are replaced by LSA similarity values that have been re-scaled to improve numerical comparability across topics. The main difference of this model and the standard supervised HMM is that the proposed model does not require labeled training data. Instead, it only requires a set of keywords defining the topics and large-enough body of unannotated text on which the LSA is calculated. The model is decoded using the standard Viterbi algorithm.

5 Performance evaluation

We evaluate the proposed method on manually annotated gold-standard data (see Section 3). The test set consists of 204 and the training set of 198 annotated shifts randomly selected from 135 patient reports. If two shifts report on the same patient, both are placed either in the train set or in the test set. LSA is calculated from all text available in the 448 patient reports from which no shift was selected into the test set.

To reduce sparseness problems due to the highly-inflective nature of Finnish, we lemmatize the text using a version of the FinTWOL Finnish morphological analyzer[1] (Koskenniemi, 1983) whose lexicon has been extended by approximately 3500 clinical domain terms. For every word analyzed by FinTWOL, we use the first lemma given, and for words outside of FinTWOL lexicon, we use the unchanged surface word form. The LSA similarity scores are computed using the Infomap NLP software[2] (Dorow and Widdows, 2003).

Since a fully-unsupervised parameter-selection method is so-far not available, we select the parameters by grid search on a held-out set of 60 annotated shifts. These shifts are not part of the test set in order to avoid overfitting the parameter selection. The context window width is set to 30 words (left and right context both 15 words), and the method parameters are $\delta = 0.6$, $\alpha = 0.3$, and $\beta = 0.15$. All other LSA-related parameters (max number of singular values, number of Word Space columns, etc.) are left at their default after preliminary experiments indicated that they have only marginal effect on the overall performance.

To establish the relative merit of the unsupervised method, we compared its performance against two other methods: a keyword-trigger method and a comparable supervised learning method. The keyword method is a simple baseline that performs segmentation and labeling by looking for the occurrence of the five topic keywords (*breathing* etc.), assigning each word to a labeled segment corresponding to the previous seen keyword. The assigned label is given the initial value *other* at the start of each shift. To allow the keyword-based approach to benefit from the normalizing effect of morphological analysis, the trigger words are matched against the lemmas given by FinTWOL.

The supervised method compared to is a basic first-order HMM. This choice is made not out of ignorance of advances such as conditional random fields (see, e.g., (Sutton et al., 2007)), but rather as HMM is a close supervised equivalent of the proposed model — we sought to determine

[1] `http://www.lingsoft.fi/`

[2] `http://infomap-nlp.sourceforge.net/`

	Accuracy	WindowDiff
majority baseline	23.4	0.32
keyword baseline	66.9	0.16
unsupervised model	74.9	0.23
supervised HMM	82.9	0.21

Table 1: Performance of the three compared methods. Note that for WindowDiff lower value indicates better performance — a perfect segmentation obtains WindowDiff score of zero. Majority baseline refers to assigning the most common topic in the data (*consciousness*) to all tokens.

the relative efficiency of the unsupervised and supervised alternatives in setting the parameters of the graphical model. For the HMM, the only parameter, the smoothing model and its setting, was selected on the training set by a separate search of the parameter space so as to avoid overfitting the test set. The selected optimal smoothing model was Lidstone (add-γ) smoothing with $\gamma = 0.3$.

The primary evaluation measure is micro-averaged accuracy, the proportion of words in the test set with correctly identified label. Further, we report macro-averaged WindowDiff (Pevzner and Hearst, 2002) score, which is often used to evaluate segmentation quality independently of the topic labels. The WindowDiff window size was set to half of the average segment size in the gold standard data, a standard way to set this parameter. Note that WindowDiff only takes into account the positions of segment boundaries, ignoring the topic labels.

6 Results and discussion

The performance of the methods on the test set (204 shifts, 15839 tokens) is reported in Table 1. As expected, the accuracy of the unsupervised model is between the performance of the keyword baseline and the supervised HMM. The unsupervised model considerably outperforms the keyword baseline. Further, it is not surprising that the supervised HMM performs better than the unsupervised model, considering that it receives much more detailed information about the distribution of words with respect to topics.

Interestingly, the WindowDiff results are in disagreement with the accuracy results, with the keyword baseline reaching better WindowDiff performance than even the supervised HMM. We have currently no explanation for this highly unintuitive secondary result. Nevertheless, as the unsupervised method performs nearly at the level of the supervised in terms of WindowDiff and this measure does not take into account the assigned labels, a key aspect of the method, we do not view this result as compromising the positive primary findings in terms of accuracy. In the following, we will only focus on accuracy results.

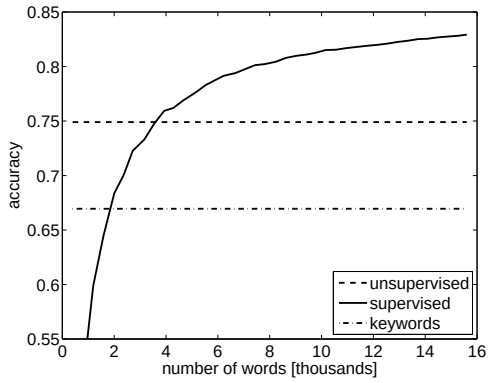

Figure 4: Learning curve for the supervised baseline method. The performance of the unsupervised and keyword-based methods are shown for reference.

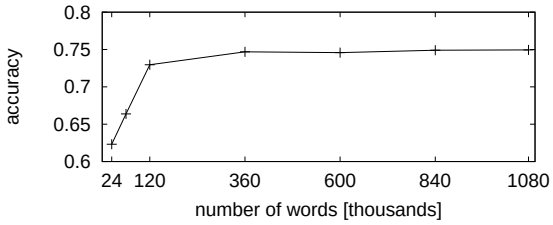

Figure 5: Learning curve for the unsupervised method.

An interesting question to ask here is how many words of labeled data does the HMM require to reach the accuracy of the unsupervised method. The learning curve of the HMM, that is the dependence of its accuracy on the amount of available training data, is given in Figure 4. Here we observe that in order to reach the performance of the unsupervised method, it is necessary to manually label roughly 3600 words. For comparison, the learning curve for the unsupervised method is shown in Figure 5; the curve is generated by varying the amount of text available to calculate the LSA. Here we see that the peak performance is reached after about 360,000 words (150 full patient reports). Note that for the unsupervised method the text is not manually labeled; gathering the amount of data necessary for reaching the peak performance does not involve any manual annotation effort, unlike in the case of the supervised HMM.

7 Conclusions and future work

We have introduced an unsupervised method for TS and labeling based on a combination HMMs and LSA. We have shown that, in order to reach the performance of the unsupervised method, a standard HMM would require 3600 words of labeled training data, as opposed to just one keyword per topic necessary for the unsupervised method. The proposed method is thus applicable to information search tasks with freely-chosen topics and no labeled data available. We have applied the method to a real-life clinical task.

In further research, several crucial questions will be investigated. First is that of unsupervised selection of the parameters of the system (such as the LSA window width and self-transition probability δ). The second open question is whether the current proposed model can be re-normalized to obtain an actual HMM without loss of performance. This would open further interesting directions such as the possibility to use the LSA-based HMM model as an initial state for further unsupervised training of the method, for instance by the standard Baum-Welch algorithm. Finally, a general way of modeling the topic *other* is needed for applications where some segments do not belong to any keyword-defined topic.

Acknowledgments

This work was supported by the Academy of Finland and the Finnish Funding Agency for Technology and Innovation, Tekes. We thank Simo Vihjanen and Sari Ahonen from Lingsoft Inc. for extending the FinTWOL lexicon and Heljä Lundgrén-Laine and Päivi Haltia for their advise regarding ICU practices and language.

References

D Beeferman, A Berger, and J Lafferty. 1997. Text segmentation using exponential models. In *Proceedings of EMNLP-2*, pages 35–46. ACL.

Y Bestgen. 2006. Improving text segmentation using Latent Semantic Analysis: A reanalysis of Choi, Wiemer-Hastings, and Moore (2001). *Computational Linguistics*, 32(1):5–12.

DM Blei and PJ Moreno. 2001. Topic segmentation with an aspect hidden Markov model. In *Proceedings of SIGIR'01*, pages 343–348. ACM.

P Bramsen, P Deshpande, YK Lee, and R Barzilay. 2006. Finding temporal order in discharge summaries. In *AMIA Annu Symp Proc 2006*, pages 81–85. AMIA.

T-H Chang and Ch-H Lee. 2003. Topic segmentation for short texts. In *Proceedings of PACLIC 17*, pages 159–165. Colips Publications.

PS Cho, RK Taira, and H Kangarloo. 2003. Automatic section segmentation of medical reports. In *AMIA Annu Symp Proc 2003*, pages 155–159. AMIA.

B Dorow and D Widdows. 2003. Discovering corpus-specific word senses. In *Proceedings of EACL'03*, pages 79–82. ACL.

O Ferret. 2002. Using collocations for topic segmentation and link detection. In *Proceedings of COLING'02*, pages 1–7. ACL.

MA Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

M Hiissa, T Pahikkala, H Suominen, T Lehtikunnas, B Back, H Karsten, S Salanterä, and T Salakoski. 2007. Towards automated classification of intensive care nursing narratives. *Int J Med Inform*, 76(S3):362–368.

K Koskenniemi. 1983. Two-level model for morphological analysis. In *Proceedings of IJCAI'83*, pages 683–685. Morgan Kaufmann.

L Pevzner and MA Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

JM Ponte and WB Croft. 1997. Text segmentation by topic. In *Proceedings of ECDL '97*, pages 113–125. Springer-Verlag.

LR Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

JC Reynar. 1999. Statistical models for topic segmentation. In *Proceedings of ACL'99*, pages 357–364. ACL.

H Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

H Suominen, S Pyysalo, F Ginter, and T Salakoski. 2008. Automated text segmentation and topic labeling of clinical narratives. In *Proceedings of Louhi'08*. TUCS. To appear.

C Sutton, A McCallum, and K Rohanimanesh. 2007. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *J Mach Learn Res*, 8:693–723.

D Widdows and S Peters. 2003. Word vectors and quantum logic: Experiments with negation and disjunction. In *Proceedings of MoL8*, pages 141–154.

JP Yamron, I Carp, L Gillick, S Lowe, and P van Mulbregt. 1998. A hidden Markov model approach to text segmentation and event tracking. In *Proceedings of ICASSP'98*, pages 333–336. IEEE.