INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS XXX (2009) XXX–XXX

ELSEVIER

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application

Filip Ginter[a,*], *Hanna Suominen*[a,b], *Sampo Pyysalo*[b], *Tapio Salakoski*[a,b]

[a] *Department of Information Technology, University of Turku, Joukahaisenkatu 3-5, 20520 Turku, Finland*[1]
[b] *Turku Centre for Computer Science (TUCS), Joukahaisenkatu 3-5, 20520 Turku, Finland*

ARTICLE INFO

Article history:
Received 31 October 2008
Received in revised form
17 December 2008
Accepted 5 February 2009

Keywords:
Hidden Markov models
Latent semantic analysis
Topic segmentation
Topic classification
Information retrieval
Computerized patient records
Nursing

ABSTRACT

Motivation: 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.
Methods: We introduce an unsupervised method based on a combination of hidden Markov models and latent semantic analysis which allows the topics of interest to be defined freely, without the need for data annotation, and can identify short segments.
Results: The method is evaluated on intensive care nursing narratives and motivated by information needs in this domain. The method 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

Topic segmentation (TS) and labeling systems enable fine-grained information search. We have previously applied a TS and labeling method, a supervised hidden Markov model (HMM), to Finnish intensive care unit (ICU) nursing narratives [1]. The problem was to automatically divide text into topically coherent segments with respect to pre-determined, repeatedly discussed topics to support information access and clinical decision-making (see Fig. 1). This type of structure has been empirically shown to increase the information search speed of clinicians [2]. However, this approach requires topic-labeled training data to induce the HMM model and consequently, TS topics cannot be changed without additional annotation effort from the perspective of the new information need.

In this paper, we introduce a TS and labeling method, where the topics are not fixed in advance but are provided by the user as freely chosen keywords (e.g., *breathing* or *hemodynamics*). It combines latent semantic analysis (LSA) with a graphical model closely related to HMMs and is unsupervised in the sense of not requiring labeled training data. This allows the topics of interest to be easily changed: the user simply specifies new keywords.

* *Corresponding author*. Tel.: +358 50 4138305.
E-mail addresses: filip.ginter@utu.fi (F. Ginter), hanna.suominen@utu.fi (H. Suominen), tapio.salakoski@utu.fi (T. Salakoski).
[1] first.last@utu.fi.
1386-5056/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2009.02.003

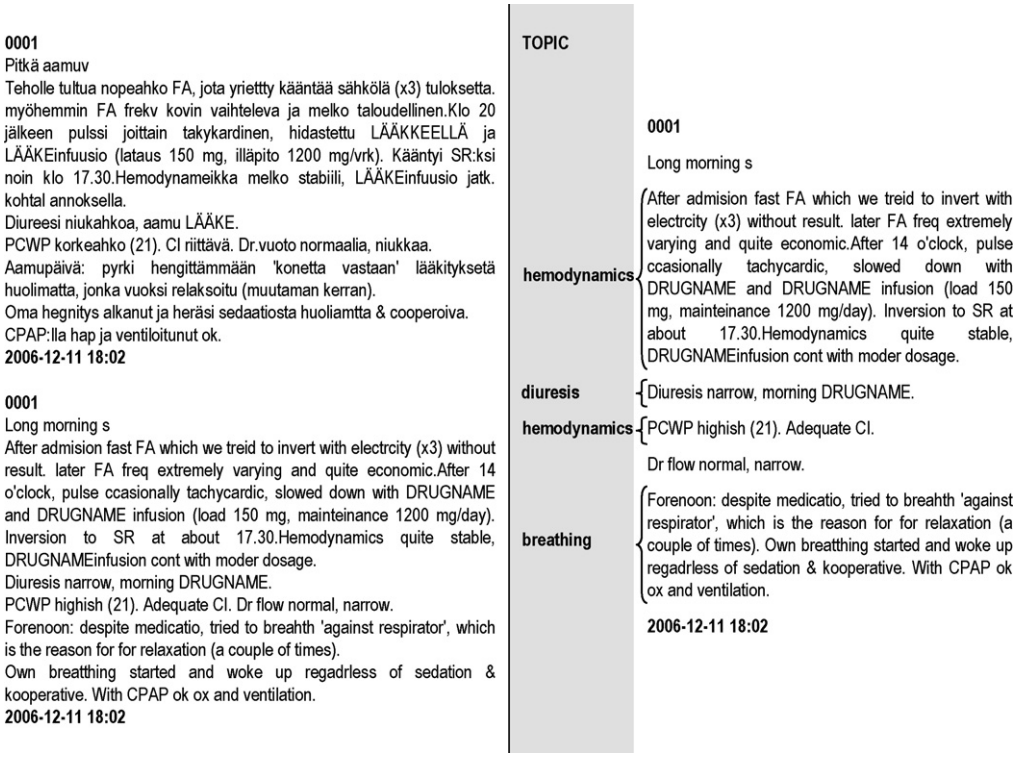

Fig. 1 – An anonymized illustration of the Finnish data translated to English preserving typographical errors and its example segmentation into topics.

The applicability of existing TS methods is limited in our case. First, to allow a free, *ad hoc* choice of topics, we require an unsupervised approach. Commonly a TS problem is solved in an unsupervised manner by analyzing the similarity (e.g., first uses of words, word co-occurrence, repetition or semantic relations) of text before and after a proposed segment boundary (see, e.g. [3,4]); a sudden drop in similarity indicates a likely change in topic. However, these techniques do not typically allow the topics of interest to be specified in advance and methods that consider pre-specified topics (see, e.g. [5–7]) tend to be supervised. Second, the ICU narratives are characterized by extremely short segments; a single sentence may contain several topic-changes and the average topic length is only 18 tokens. Existing unsupervised TS methods require considerably longer segments (e.g., the TextTiling method [3] searches for topic boundaries between contexts of 200 tokens) and those specifically designed for short segments (see, e.g. [8,9]) do not consider pre-specified topics. Third, our method is specifically designed for applications where almost all documents contain relevant information about the topics of interest.

Although ICU narratives motivated us for developing the method, it is a general TS and labeling technique that we believe to have potential to support *ad hoc* information needs also in many other application domains. For instance, the method could be applicable in more general information retrieval tasks such as rhetorical zone detection [10].

2. Clinical data

The dataset used in this study consists of nursing notes of 516 adult ICU patients.[2] These Finnish patient-specific records are written during every shift mainly for intra-unit information exchange. The dataset consists of 17,140 nursing shifts.

We apply a simple domain-adapted tokenizer, obtaining 1.2 million tokens (including punctuation). The most common topics of the text are *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. The text can be characterized as telegraphic and highly specialized with a substantial amount of professional terminology, unit-specific documentation practices, and misspellings (Fig. 1).

For testing, we use a subset consisting of 402 shifts randomly chosen from the records of the first 135 patients by their admission date [1]. This test dataset is manually segmented and labeled by a domain expert with respect to the topics listed above using the Knowtator tool [11] of Protégé 3.3.1 Ontology Editor and Knowledge Acquisition System[3]; text not belonging to any of these is assigned the topic *other*.

[2] Collected retrospectively from January 1, 2005 to August 1, 2006 with proper permissions (Statutes of Finland: Medical research act 488/1999 and decree 986/1999).

[3] http://protege.stanford.edu/.

Please cite this article in press as: F. Ginter, et al., Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application, Int. J. Med. Inform. (2009), doi:10.1016/j.ijmedinf.2009.02.003

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

3.1. 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. While the standard vector-space model represents documents in a sparse, high-dimensional space where each dimension corresponds to a single word, LSA reduces this space into a relatively low-dimensional *latent semantic space* with the number of dimensions typically on the order of at most several hundreds. Both documents and individual words can be "folded" into this space and their similarity measured by the standard cosine metric.

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 LSA model [12] 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. The value of $lsa(w_j, q_i)$ is the cosine similarity of the vectors representing w_j and the keyword that defines q_i in the latent semantic space. That is, $lsa(w_j, q_i)$ is the commonly used LSA-based term–term similarity measure. The LSA similarity values range between 0 and 1 where larger values correspond to higher similarity.

The values of $lsa(w_j, q_i)$ 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) using the aforementioned cosine similarity. In our experiments, we use the Finnish equivalents of the keywords *breathing*, *hemodynamics*, *consciousness*, *relatives* 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 relatives NOT diuresis*. The negation operator *NOT* is available in Word Space LSA queries [13]. The resulting LSA similarity values are illustrated in Fig. 2. Punctuation, numbers, and a small number of extremely common stop-words are excluded from the LSA calculation.

3.2. 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 $w(t)$ is only dependent on the value of the hidden variable $q(t)$. Additionally, we assign a uniform distribution to the initial probability of the states. The labeling given by the HMM is the best hidden state sequence $\hat{q}$ obtained by solving

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

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

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

The optimal sequence $\hat{q}$ is known as the Viterbi path and the optimization problem (1) can be efficiently solved using the standard Viterbi algorithm. For a detailed introduction to these algorithms, see, for example, the introduction [14].

3.3. The proposed unsupervised method

In order to solve (1), the conditional probabilities $P(w(t)|q(t))$, typically referred to as *emission probabilities*, and $P(q(t)|q(t-1))$,

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

Fig. 2 – The words most similar to the topics *relatives*, *hemodynamics* and *other* together with their associated LSA similarity values.

Please cite this article in press as: F. Ginter, et al., Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application, Int. J. Med. Inform. (2009), doi:10.1016/j.ijmedinf.2009.02.003

typically referred to as *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.

3.3.1. *Transition probabilities* $P(q_j|q_i)$

Given the lack of annotated data for direct estimation of the transition probabilities, we model them with a uniform distribution. However, we need to control the segmentation granularity and thus we introduce a *self-transition probability* parameter $\delta \in (0, 1)$. The HMM transition probability is then defined as

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

The probability of continuing the current topic is thus δ, and the remaining probability $1 - \delta$ of switching a topic is distributed uniformly. While this model does not capture the natural Markov dependencies between the topics, it is fully unsupervised.

3.3.2. *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. As discussed in more detail in the study [15], we use the unnormalized LSA similarity 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

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

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. The LSA similarity values are further re-scaled to improve numerical comparability across topics as discussed in detail in our previous study [15].

4. Evaluation

We evaluate the proposed method on manually annotated data (see Section 2) randomly selected from 135 patient reports and divided among training (198 shifts) and testing (204 shifts). If two shifts report on the same patient, both are placed either in the training set or in the test set. To deal with the highly inflective nature of Finnish, we lemmatize the text using the FinTWOL Finnish morphological analyzer[4] [16] in all experiments. Our version has a lexicon extended by approximately 3500 clinical domain terms. For words analyzed by FinTWOL, we use the first lemma given, and otherwise, we use the unchanged surface word form.

LSA is performed with the Infomap NLP software[5] [17] on all text available in the 448 patient reports from which no shift was selected for testing. In the absence of a fully unsupervised parameter-selection method, we specify all parameter values by a grid search on 60 annotated shifts, finding the global maximum of performance in the parameter space. To avoid overfitting, the parameter-selection shifts are held out from testing. The resulting parameter values are: left and right context window width of 15 words, $\delta = 0.6$, $\alpha = 0.3$, and $\beta = 0.15$, where α and β are scaling parameters (see [15]). For all other LSA-related parameters (maximal number of singular values, number of Word Space columns, etc.), we use the default values, as our preliminary experiments indicated they only have a marginal effect on the overall performance.

To establish the relative merit of the unsupervised method, we compare its performance against a close supervised equivalent [1] and a topic keyword search inherently resembling the structure of our data. The latter method searches for the occurrence of the five topic keywords (*breathing*, etc.) and assigns 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.

As an evaluation measure, we use primarily micro-averaged accuracy but we report also macro-averaged WindowDiff [18], which evaluates TS quality independently of the topic labels. The WindowDiff window size was set to half of the average segment size in the manually annotated data, a standard way to set this parameter.

Table 1 – Performance of the compared methods on the test set (204 shifts, 15,839 tokens). Majority baseline refers to assigning always the most common topic label (*consciousness*). For WindowDiff a lower value indicates better performance.

	Accuracy	WindowDiff
Majority baseline	23.4	0.32
Keyword search	66.9	0.16
Unsupervised model	74.9	0.23
Supervised HMM	82.9	0.21

5. Results

The accuracy of the unsupervised model is considerably better than that of the keyword baseline, but, as expected, it is outperformed by the supervised HMM as the latter receives much more detailed information about the distribution of words with respect to topics (Table 1). To reach the performance of the unsupervised method, the supervised HMM requires approximately 3600 words of manually labeled training data (Fig. 3). For comparison, the learning

[4] http://www.lingsoft.fi/.
[5] http://infomap-nlp.sourceforge.net/.

Please cite this article in press as: F. Ginter, et al., Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application, Int. J. Med. Inform. (2009), doi:10.1016/j.ijmedinf.2009.02.003

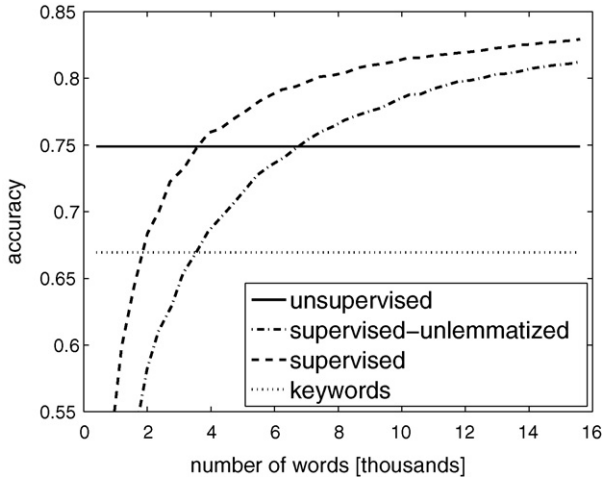

Fig. 3 – Learning curve for the supervised baseline method with and without the use of FinTWOL. The performance of the unsupervised and keyword-based methods are shown for reference.

curve for the unsupervised method is shown in Fig. 4. We see that the peak performance is reached after about 360,000 words (150 full patient reports). Note that for the unsupervised method 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.

Unintuitively, the WindowDiff results (Table 1) are in disagreement with the accuracy results, with the keyword baseline reaching better WindowDiff performance than even the supervised HMM. 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.

To evaluate the effect of linguistic processing, we consider the learning curves of the supervised HMM on datasets with and without lemmatization (Fig. 3). Lemmatization improves the performance, but, as would be intuitively expected, its significance diminishes with increasing amount of training data.

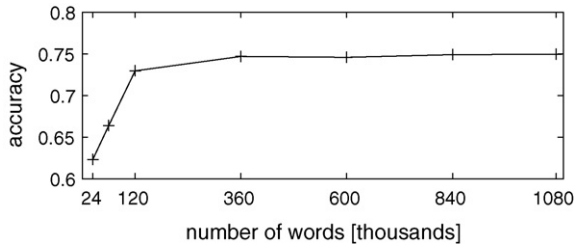

Fig. 4 – Learning curve for the unsupervised method. The curve is generated by varying the amount of text available to calculate the LSA.

6. Conclusions and future work

We have introduced an unsupervised method for TS and labeling based on a combination of HMMs and LSA and applied the proposed method in a clinically motivated setting. 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.

Our study holds promise for improving the functionality of electronic patient records. A topic-wise highlighting system is likely to improve the capabilities for searching information from narratives, building trends in time and summarizing the gathered notes. The proposed unsupervised method is applicable to information search tasks with freely chosen topics and no labeled data available. However, if the search topics are established and resources for manual labeling exists, supervised methods are preferred because they offer higher performance.

In further research, we focus on unsupervised selection of the parameters of the system (e.g., the LSA window width and self-transition probability δ). Another topic of further work is to study the effect of various other methods providing unsupervised similarity measures such as the probabilistic LSA [19] and Random Indexing [20]. Additionally, the use of the probabilistic interpretation of LSA might lead to a proper HMM, which would open further interesting directions. For example, the LSA-based HMM model could be used as an initial state for further unsupervised training of the method with 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.

A pilot study will be performed to evaluate the benefits of the method in clinical practice. Further, the applicability of the method as a pre-analysis step in trend building and summarization systems will be evaluated.

Summary points

What was known before:

- Topic segmentation and labeling systems enable fine-grained information search.
- Best performance is achieved by supervised methods, which, however, do not allow *ad hoc* information search.
- There is a need for *ad hoc* information search in clinical narratives.

What this study added to the body of knowledge:

- A novel method to induce an unsupervised text segmentation model where the topics of interest are specified as keywords with no need for training data annotation.
- The unsupervised method notably outperforms a baseline based on keyword triggers.
- A corresponding supervised method needs at least 3600 manually annotated words to reach the performance of the

Please cite this article in press as: F. Ginter, et al., Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application, Int. J. Med. Inform. (2009), doi:10.1016/j.ijmedinf.2009.02.003

unsupervised method, rendering the supervised method inapplicable in *ad hoc* segmentation.
- Lemmatization contributes to the performance, mitigating the effects of data sparseness due to inflections.

Acknowledgments

This work was supported by the Academy of Finland and the Finnish Funding Agency for Technology and Innovation, Tekes (40020/07). We thank Sari Ahonen and Simo Vihjanen from Lingsoft Inc. for extending FinTWOL, Philip Ogren for assistance with Knowtator and Heljä Lundgrén-Laine for help in annotation.

REFERENCES

[1] H. Suominen, S. Pyysalo, F. Ginter, T. Salakoski, Automated text segmentation and topic labeling of clinical narratives, in: Proceedings of Louhi'08, TUCS, 2008, pp. 99–103.
[2] H.J. Tange, H.C. Schouten, A.D.M. Kester, A. Hasman, The granularity of medical narratives and its effect on the speed and completeness of information retrieval, Journal of American Medical Informatics Association 5 (6) (1998) 571–582.
[3] M.A. Hearst, TextTiling: segmenting text into multi-paragraph subtopic passages, Computational Linguistics 23 (1) (1997) 33–64.
[4] O. Ferret, Using collocations for topic segmentation and link detection, in: Proceedings of COLING'02, ACL, 2002, pp. 1–7.
[5] J.P. Yamron, I. Carp, L. Gillick, S. Lowe, P.van Mulbregt, A hidden Markov model approach to text segmentation and event tracking, in: Proceedings of ICASSP'98, IEEE, 1998, pp. 333–336.
[6] D.M. Blei, P.J. Moreno, Topic segmentation with an aspect hidden Markov model, in: Proceedings of SIGIR'01, ACM, 2001, pp. 343–348.
[7] A. Gruber, M. Rosen-Zvi, Y. Weiss, Hidden topic Markov models, in: Proceedings of AISTATS'07, Society for Artificial Intelligence and Statistics, 2007.
[8] J.M. Ponte, W.B. Croft, Text segmentation by topic, in: Proceedings of ECDL'97, Springer-Verlag, 1997, pp. 113–125.
[9] T.-H. Chang, Ch.-H. Lee, Topic segmentation for short texts, in: Proceedings of PACLIC 17, Colips Publications, 2003, pp. 159–165.
[10] T. Mullen, Y. Mizuta, N. Collier, A baseline feature set for learning rhetorical zones using full articles in the biomedical domain, SIGKDD Explorations 7 (1) (2005) 52–58.
[11] P.V. Ogren, Knowtator: a Protégé plug-in for annotated corpus construction, in: Proceedings of HLT-NAACL'06, ACL, 2006, pp. 273–275.
[12] H. Schütze, Automatic word sense discrimination, Computational Linguistics 24 (1) (1998) 97–123.
[13] D. Widdows, S. Peters, Word vectors and quantum logic: experiments with negation and disjunction, in: Proceedings of MoL8, 2003, pp. 141–154.
[14] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.
[15] F. Ginter, H. Suominen, S. Pyysalo, T. Salakoski, Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: method and clinical application, in: Proceedings of SMBM'08, TUCS, 2008, pp. 37–44.
[16] K. Koskenniemi, Two-level model for morphological analysis, in: Proceedings of IJCAI'83, Morgan Kaufmann, 1983, pp. 683–685.
[17] B. Dorow, D. Widdows, Discovering corpus-specific word senses, in: Proceedings of EACL'03, ACL, 2003, pp. 79–82.
[18] L. Pevzner, M.A. Hearst, A critique and improvement of an evaluation metric for text segmentation, Computational Linguistics 28 (1) (2002) 19–36.
[19] T. Hofmann, Probabilistic latent semantic analysis, in: Proceedings of UAI'99, Morgan Kaufmann, 1999, pp. 289–296.
[20] P. Kanerva, J. Kristofersson, A. Holst, Random indexing of text samples for latent semantic analysis, in: Proceedings of CogSci'00, Erlbaum, 2000, p. 1036.

Please cite this article in press as: F. Ginter, et al., Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application, Int. J. Med. Inform. (2009), doi:10.1016/j.ijmedinf.2009.02.003