Automated Text Segmentation and Topic Labeling of Clinical Narratives

Hanna Suominen[1,2], Sampo Pyysalo[1,2], Filip Ginter[2], and Tapio Salakoski[1,2]

[1] Turku Centre for Computer Science (TUCS) and
[2] University of Turku, Department of Information Technology
Joukahaisenkatu 3-5 B, 20520 Turku, Finland
firstname.lastname@utu.fi

Abstract. Electronic patient information systems include numerous functionalities to support clinical judgment and decision-making, but their capabilities to analyze free-text narratives are limited. We apply Hidden Markov Models to divide Finnish intensive care nursing notes into topically coherent segments and assign a topic label to each segment. The method notably outperforms a keyword-based baseline already with a relatively small amount of training data. The result holds the promise of increased information search speed and a more comprehensive overall picture about patients.

1 Introduction

Modern electronic patient information systems include numerous functionalities to support clinical judgment and decision-making [1]. For example, they generate statistics, trends and alerts from the patient data. However, although a substantial amount of information is documented as free-text notes, referred to as narratives, automated processing is typically limited to the numerical or structured parts of the patient records. Text mining applications in clinical use, such as MedLEE [2] and Autocoder [3] for English text, are rare in particular for minority languages.

In this study, we present a text segmentation and topic labeling (TS & TL) method for dividing Finnish narratives automatically into topically coherent, non-overlapping segments (Figure 1). The resulting type of structure has been empirically shown to increase the information search speed of clinicians [4]. The domain we consider is intensive care (IC), as its complexity, information richness, and fast pace make decision-making particularly challenging.

In the clinical domain, TS techniques have previously been applied, for example, to temporal order analysis of medical discharge summaries [5]. The method solves the problem sequentially by using a statistical parser to segment the sentences into clauses, a classifier to predict the segment boundaries between the clauses and finally another classifier to decide for every segment pair their timewise order. Another application related to this study is TS & TL of medical narratives from radiology and urology departments [6]. Although it contains

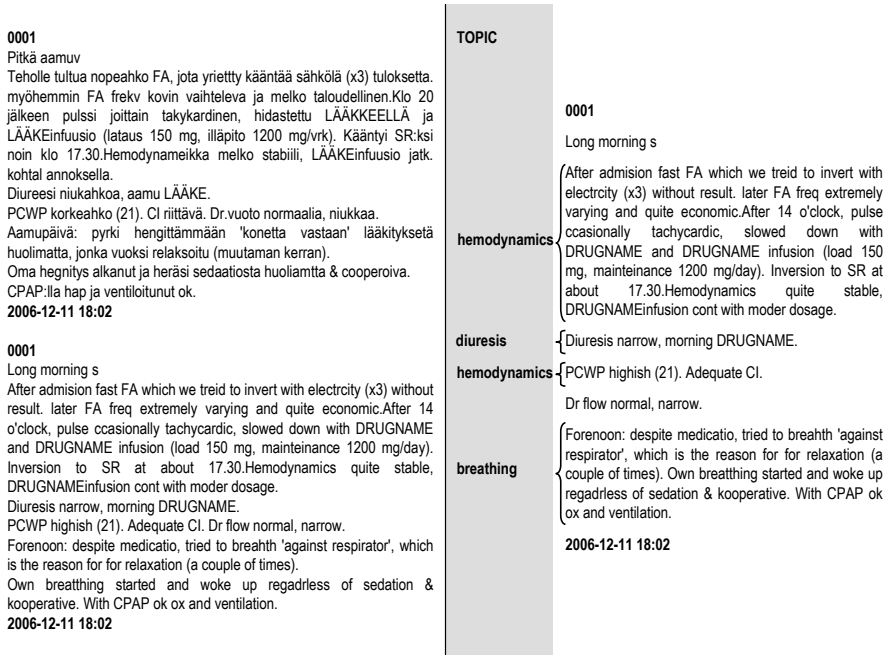

Fig. 1. An anonymized illustration of the Finnish data accompanied with its English translation preserving typographical errors and an example segmentation into topics.

a classifier based on the order of sections and other statistical features of the training data, it mainly relies on hard-coded headlining rules, linguistic cues and lexical patterns seen within training examples. Finally, a system classifying segments of IC patient narratives with respect to the topics of *breathing*, *blood circulation* and *pain* has been introduced [7]; it does not, however, perform automated TS at all.

2 Patient data and its linguistic processing

Anonymized nursing narratives[3] of 516 adult IC patients were used in this study. They covered the whole in-patient time and were written mainly for intra-unit information exchange. We chose these notes because their use in direct care is hindered due to the large quantity of text.

The data set included altogether 17140 patient and nursing shift-specific documents, which we call shifts (Figure 1). Each shift contained, on average, 73 tokens (including punctuation). The vocabulary was highly specialized, with a

[3] 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) for the Louhi project (www.med.utu.fi/hoitotiede/tutkimus/tutkimusprojektit/louhi).

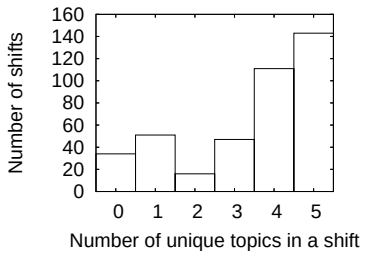

Fig. 2. A histogram illustrating the number of topics discussed per shift.

substantial amount of unit-specific practices and domain terminology. Approximately half of the shifts were structured by using colon-separated, but non-standardized, headings, as *Hemodynamics*, *HAEMODYNAMICS*, *H e m o d*, and *Homedynamics*. The most common documentation topics were *breathing*, *hemodynamics*, *consciousness*, *relatives*, and *diuresis*. We selected these to be used as TS topics due to their prevalence in the narratives; by recognizing them automatically, building an overall picture about their development in time could be supported, for example, through topical highlighting.

To create topic-annotated data for experiments, we randomly chose three shifts per patient from the records of 135 patients and tagged text segments relevant to the five most common topics by using the Knowtator tool [8] of Protégé 3.3.1 Ontology Editor and Knowledge Acquisition System[4]. Irrelevant parts were given the label *other*. With problematic phrases, an IC nursing specialist was consulted. The average shift length was 78 tokens while the average segment length was only 18 tokens. Typically, all or almost all five topics were discussed within one shift although 34 shifts contained none of the topics (Figure 2).

To reduce data sparseness caused by the highly inflective nature of Finnish, we lemmatized the data using a version of the FinTWOL Finnish morphological analyser[5] [9] whose lexicon was extended by approximately 3500 clinical domain terms. For every word analyzed by FinTWOL, we used the first lemma given.

3 Method and its performance evaluation setting

Let us denote the topics of interest as $q_i, i \in \{1, \ldots, N_q\}$. Our TS & TL task is to infer for the input word sequence $w = [w(1) \ldots w(T)]$ the topic sequence $q = [q(1) \ldots q(T)]$, where $w(t)$ belongs to the vocabulary $\{w_1, \ldots, w_{N_w}\}$ of N_w unique words and $q(t) \in \{q_1, \ldots, q_{N_q}\}$ for all $t \in \{1, \ldots, T\}$. A convenient way to model this sequence labeling problem is to use a first-order Hidden Markov Model (HMM) (see, e.g., [10]), where a particular hidden variable $q(t)$ only depends on the previous hidden state $q(t-1)$, an observed variable $w(t)$ is only dependent on the value of the hidden variable $q(t)$, and the random variable

[4] http://protege.stanford.edu/
[5] http://www.lingsoft.fi/

describing the start of the chain is uniformly distributed. Formally, if $\mathcal{Q}$ is the space of all hidden state sequences, we infer the best q by solving

$$\arg\max_{q \in \mathcal{Q}} P(w(1)|q(1)) \prod_{t=2}^{T} P(w(t)|q(t)) P(q(t)|q(t-1)).$$

We trained the HMM with approximately half of the annotated shifts and tested it with the other half. No patient record was divided between the two sets. The smoothing model and its parameter were selected on the training set by a separate search of the parameter space so as to avoid over-fitting the test set. The selected optimal model was Lidstone (add-γ) smoothing (see, e.g., [11, p. 204]) with $\gamma = 0.3$.

The baseline algorithm implements a simple topic keyword search: We first searched for the five topic keywords. Then, we assigned each word to a labeled segment corresponding to the previous seen topic until the end of the shift. We gave the assigned label at the start of each shift the initial value of *other*. This baseline was chosen because it inherently resembles the documentation structure. To allow the baseline to benefit from the normalizing effect of morphological analysis, TS & TL was performed with the data processed with FinTWOL. In evaluation, we measured the token-wise average TL accuracy over the whole test data.

4 Results and conclusion

HMM performing TS & TL of IC nursing notes with the lemmatized text notably outperformed the keyword-based baseline of 66.95% already with as few as 2000 words of training data (Figure 3). This corresponds approximately to tagging 20 shifts like the one given in Figure 1. The performance increase can be seen to level off after about 8000 words. Linguistic processing contributed to the performance, but its significance diminished by increasing the amount of training data: the accuracy of HMM with the lemmatized data was 82.93%, whereas the respective number without lemmatization was 81.22%.

Our results hold promise for improving the functionality of electronic patient information systems: the method is easy to implement and its integration should be relatively straightforward. Highlighting of the most prevalent topics is likely to expedite information search and offer improved capabilities to build an overall picture about their development in time. In order to allow freely chosen segmentation topics, we have also developed an unsupervised method for the task [12]. Future work will include a pilot study testing our methods in clinical use. Other interesting research directions are generating trends and summarizing text on the basis of the automatically topic-labeled narratives.

Acknowledgments

We gratefully acknowledge the financial support of the Academy of Finland and the Finnish Funding Agency for Technology and Innovation, Tekes. We also express our

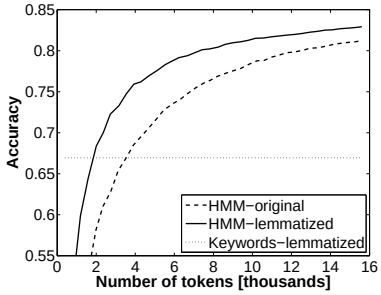

Fig. 3. Learning curves for the HMMs with and without linguistic processing and the overall accuracy of the keyword-based baseline.

gratitude to Heljä Lundgrén-Laine for help in annotation, to Philip Ogren for assistance with Knowtator and to Sari Ahonen and Simo Vihjanen from Lingsoft Inc. for extending FinTWOL.

References

1. Hanson, C., Marshall, B.: Artificial intelligence applications in the intensive care unit. Crit Care Med **29**(2) (2001) 427–435
2. Mendonça, E., Haas, J., Shagina, L., Larson, E., Friedman, C.: Extracting information on pneumonia in infants using natural language processing of radiology reports. J Biomed Inform **38**(4) (2005) 314–321
3. Pakhomov, S., Buntrock, J., Chute, C.: Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. J Am Med Inform Assoc **13**(5) (2007) 516–525
4. Tange, H., Schouten, H., Kester, A., Hasman, A.: The granularity of medical narratives and its effect on the speed and completeness of information retrieval. J Am Med Inform Assoc **5**(6) (1998) 571–582
5. Bramsen, P., Deshpande, P., Lee, Y., Barzilay, R.: Finding temporal order in discharge summaries. AMIA Annu Symp Proc (2006) 81–85
6. Cho, P., Taira, R., Kangarloo, H.: Automatic section segmentation of medical reports. AMIA Annu Symp Proc (2003) 155–159
7. Hiissa, M., Pahikkala, T., Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salanterä, S., Salakoski, T.: Towards automated classification of intensive care nursing narratives. Int J Med Inform **76**(S3) (2007) S362–S368
8. Ogren, P.: Knowtator: A Protégé plug-in for annotated corpus construction. In: Proc HLT-NAACL 2006, Morristown, NJ, USA, ACL (2006) 273–275
9. Koskenniemi, K.: Two-level model for morphological analysis. In: Proc IJCAI 83. Volume 2., Karlsruhe, Germany, Morgan Kaufmann (1983) 683–685
10. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2) (1989) 257–286
11. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA, USA (1999)
12. Ginter, F., Suominen, H., Pyysalo, S., Salakoski, T.: Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. In: Proceedings of SMBM'08. (2008) To appear.