Österreichisches Forschungsinstitut für /
Austrian Research Institute for /
Artificial Intelligence

TR–2008–16

*Johannes Matiasek, Jeremy Jancsary,
Alexandra Klein, Harald Trost*

Identifying Segment Topics in Medical Dictations

• Freyung 6/6 • A-1010 Vienna • Austria •

• Phone: +43-1-5336112 •

• mailto:sec@ofai.at •
• http://www.ofai.at/ •

Österreichisches Forschungsinstitut für /
Austrian Research Institute for /
Artificial Intelligence

TR–2008–16

Johannes Matiasek, Jeremy Jancsary,
Alexandra Klein, Harald Trost

Identifying Segment Topics in Medical Dictations

The Austrian Research Institute for Artificial Intelligence is supported by the
Austrian Federal Ministry for Science and Research and the
Austrian Federal Ministry for Transport, Innovation and Technology.

Citation: Matiasek J., Jancsary J., Klein A., Trost H.: Identifying Segment Topics in Medical Dictations. Technical Report, Österreichisches Forschungsinstitut für Artificial Intelligence, Wien, TR-2008-16, 2008

Identifying Segment Topics in Medical Dictations

Johannes Matiasek, Jeremy Jancsary
Alexandra Klein
Austrian Research Institute for
Artificial Intelligence
Freyung 6, Wien, Austria
firstname.lastname@ofai.at

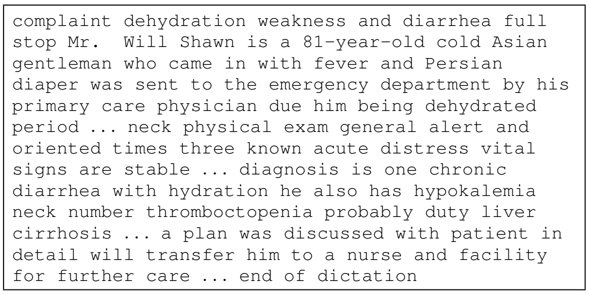
Harald Trost
Department of Medical Cybernetics
and Artificial Intelligence
of the Center for Brain Research,
Medical University Vienna, Austria
harald.trost@meduniwien.ac.at

Abstract

In this paper, we describe the use of lexical and semantic features for topic classification in dictated medical reports. First, we employ SVM classification to assign whole reports to coarse work-type categories. Afterwards, text segments and their topic are identified in the output of automatic speech recognition. This is done by assigning work-type-specific topic labels to each word based on features extracted from a sliding context window, again using SVM classification utilizing semantic features. Classifier stacking is then used for a posteriori error correction, yielding a further improvement in classification accuracy.

1 Introduction

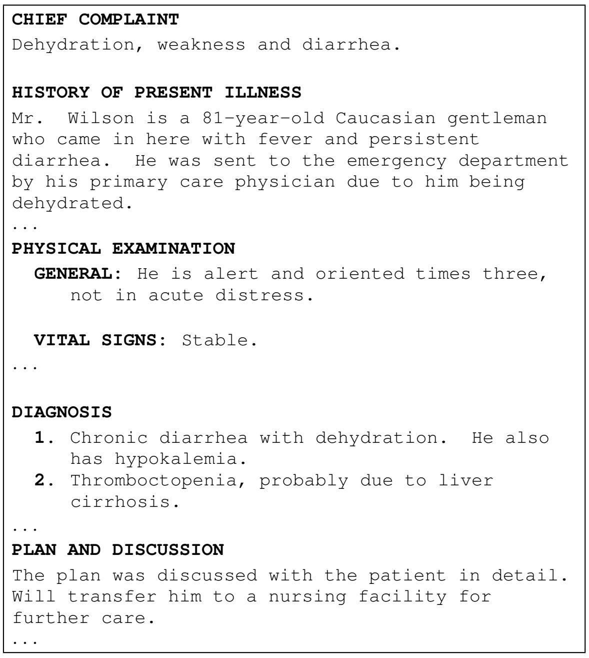
The use of automatic speech recognition (ASR) is quite common in the medical domain, where for every consultation or medical treatment a written report has to be produced. Usually, these reports are dictated and transcribed afterwards. The use of ASR can, thereby, significantly reduce the typing efforts, but, as can be seen in figure 1, quite some work is left.



```
complaint dehydration weakness and diarrhea full
stop Mr.  Will Shawn is a 81-year-old cold Asian
gentleman who came in with fever and Persian
diaper was sent to the emergency department by his
primary care physician due him being dehydrated
period ... neck physical exam general alert and
oriented times three known acute distress vital
signs are stable ... diagnosis is one chronic
diarrhea with hydration he also has hypokalemia
neck number thromboctopenia probably duty liver
cirrhosis ... a plan was discussed with patient in
detail will transfer him to a nurse and facility
for further care ... end of dictation
```

Figure 1: Raw output of speech recognition

When properly edited and formatted, the same dictation appears significantly more comprehensible, as can be seen in figure 2.



```
CHIEF COMPLAINT
Dehydration, weakness and diarrhea.

HISTORY OF PRESENT ILLNESS
Mr.  Wilson is a 81-year-old Caucasian gentleman
who came in here with fever and persistent
diarrhea.  He was sent to the emergency department
by his primary care physician due to him being
dehydrated.
...
PHYSICAL EXAMINATION
  GENERAL: He is alert and oriented times three,
     not in acute distress.

  VITAL SIGNS: Stable.
...

DIAGNOSIS
  1. Chronic diarrhea with dehydration.  He also
     has hypokalemia.
  2. Thromboctopenia, probably due to liver
     cirrhosis.
...
PLAN AND DISCUSSION
The plan was discussed with the patient in detail.
Will transfer him to a nursing facility for
further care.
...
```

Figure 2: A typical medical report

Besides the usual problem with recognition errors, section headers are often not dictated or hard to recognize as such. One task that has to be performed in order to arrive at the structured report shown in figure 2 is therefore to identify topical sections in the text and to classify them accordingly.

In the following, we first describe the problem setup, the steps needed for data preparation, and the division of the classification task into subproblems. We then describe the experiments performed and their results.

In the outlook we hint at ways to integrate this approach with another, multilevel, segmentation framework.

2 Data Description and Problem Setup

Available corpus data consists of raw recognition results and manually formatted and corrected reports of medical dictations. 11462 reports were available in both forms, 51382 reports only as corrected transcripts. When analysing the data, it became clear that the structure of segment topics varied strongly across different work-types. Thus we decided to pursue a two-step approach: firstly classify reports according to their work-type and, secondly, train and apply work-type specific classification models for segment topic classification.

2.1 Classification framework

For all classification tasks discussed here, we employed support-vector machines (SVM, Vapnik (1995)) as the statistical framework, though in different incarnations and setups. SVMs have proven to be an effective means for text categorization (Joachims, 1998) as they are capable to robustly deal with high-dimensional, sparse feature spaces. Depending on the task, we experimented with different feature weighting schemes and SVM kernel functions as will be described in section 3.

2.2 Features used for classification

The usual approach in text categorization is to use bag-of-word features, i.e. the words occuring in a document are collected disregarding the order of their appearance. In the domain of medical dictation, however, often abbreviations or different medical terms may be used to refer to the same semantic concept. In addition, medical terms often are multi-word expressions, e.g., "coronary heart disease". Therefore, a better approach for feature mapping is needed to arrive at features at an appropriate generalization level:

- Tokenization is performed using a large finite-state lexicon including multi-word medical concepts extracted from the UMLS medical metathesaurus (Lindberg et al., 1993). Thus, multi-word terms remain intact. In addition, numeric quantities in special (spoken or written) formats or together with a dimension are mapped to semantic types (e.g. "blood pressure" or "physical quantity"), also using a finite-state transducer.

- The tokens are lemmatized and, if possible, replaced by the UMLS *semantic concept* identifier(s) they map to. Thus, "CHD", "coronary disease" and "coronary heart disease" all map to the same concept "C0010068".

- In addition, also the UMLS *semantic type*, if available, is used as a feature, so, in the example above, "B2.2.1.2.1" (Disease or Syndrome) is added.

- Since topics in a medical report roughly follow an order, for the segment topic identification task also the relative position of a word in the report (ranging from -1 to +1) is used.

We also explored different weighting schemes:

- **binary**: only the presence of a feature is indicated.

- **term frequency**: the number of occurences of a feature in the segment to be classified is used as weight.

- **TFIDF**: a measure popular from information retrieval, where $tfidf_{i,j}$ of term t_i in document $d_j \in D$ is usually defined as

$$\frac{\mathrm{ct}_{i,j}}{\sum_i \mathrm{ct}_{i,j}} . \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

An example of how this feature extraction process works is given below:

token(s)	**feature(s)**	comment
...		
an		*stop word*
78 year old	QH_OLD	*pattern-based type*
female	C0085287	*UMLS concept*
	A2.9.2	*UMLS semtype*
intubated	intubate	*lemmatized (no concept)*
with		*stop word*
lung cancer	C0242379	*UMLS concept*
	C0684249	*UMLS concept*
	B2.2.1.2.1.2	*UMLS semtype*
...		

2.3 Data Annotation

For the first classification task, i.e. work-type classification, no further annotation is necessary, every report in our data corpus had a label indicating the work-type. For the segment topic classification task, however, every token of the report had to be assigned a topic label.

2.3.1 Analysis of Corrected Transcripts

For the experiments described here, we concentrated on the "Consultations" work-type, for which clear structuring recommendations, such as E2184-02 (ASTM International, 2002), exist. However, in practice the structure of medical reports shows high variation and deviations from the guidelines, making it harder to come up with an appropriate set of class labels. Therefore, using the aforementioned standard, we assigned the headings that actually appeared in the data to the closest type, introducing new types only when absolutely necessary. Thus we arrived at 23 heading classes. Every (possibly multi-word) token was then labeled with the heading class of the last section heading occurring before it in the text using a simple parser.

2.3.2 Aligment and Label Transfer

When inspecting manually corrected reports (cf. fig. 2), one can easily identify a heading and classify the topic of the text below it accordingly. However, our goal is to develop a model for identifying and classifying segments in the *dictation*, thus we have to map the annotation of corrected reports onto the corresponding ASR output. The basic idea here is to align the tokens of the corrected report with the tokens in ASR output and to copy the annotations (cf. figure 3). There are some problems we have to take care of during alignment:

1. non-dictated items in the corrected test (e.g. punctuation, headings)
2. dictated words that do not occur in the corrected text (meta instructions, repetitions)
3. non-identical but corresponding items (recognition errors, reformulations)

For this alignment task, a standard string-edit distance based method is not sufficient. Therefore, we augment it with a more sophisticated cost function. It assigns tokens that are similar (either from a semantic or from a phonetic point of view) a low cost for substitution, whereas dissimilar tokens receive a prohibitively expensive score. Costs for deletion and insertion are assigned inversely. Semantic similarity is computed using Wordnet (Fellbaum, 1998) and UMLS. For phonetic matching, the Metaphone algorithm (Philips, 1990) was used (for details see Huber et al. (2006) and Jancsary et al. (2007)).

corrected report		OP	ASR output	
...	...	...	...	...
ChiefCompl	CHIEF	del		
ChiefCompl	COMPLAINT	sub	complaint	ChiefCompl
ChiefCompl	Dehydration	sub	dehydration	ChiefCompl
ChiefCompl	,	del		
ChiefCompl	weakness	sub	weakness	ChiefCompl
ChiefCompl	and	sub	and	ChiefCompl
ChiefCompl	diarrhea	sub	diarrhea	ChiefCompl
ChiefCompl	.	sub	fullstop	ChiefCompl
HistoryOfP	Mr.	sub	Mr.	HistoryOfP
HistoryOfP	Wilson	sub	Will	HistoryOfP
		ins	Shawn	HistoryOfP
HistoryOfP	is	sub	is	HistoryOfP
HistoryOfP	a	sub	a	HistoryOfP
HistoryOfP	81-year-old	sub	81-year-old	HistoryOfP
HistoryOfP	Caucasian	sub	cold	HistoryOfP
HistoryOfP		ins	Asian	HistoryOfP
HistoryOfP	gentleman	sub	gentleman	HistoryOfP
HistoryOfP	who	sub	who	HistoryOfP
HistoryOfP	came	sub	came	HistoryOfP
HistoryOfP	in	del		
HistoryOfP	here	sub	here	HistoryOfP
HistoryOfP	with	sub	with	HistoryOfP
HistoryOfP	fever	sub	fever	HistoryOfP
HistoryOfP	and	sub	and	HistoryOfP
HistoryOfP	persistent	sub	Persian	HistoryOfP
HistoryOfP	diarrhea	sub	diaper	HistoryOfP
HistoryOfP	.	del		
...	...	...	...	...

Figure 3: Mapping labels via alignment

3 Experiments

3.1 Work-Type Categorization

In total we had 62844 written medical reports with assigned work-type information from different hospitals, 7 work-types are distinguished. We randomly selected approximately a quarter of the reports as the training set, the rest was used for testing. The distribution of the data can be seen in table 1.

Trainingset		Testset		Work-Type
649	4.1	1966	4.2	CA Cardiology
7965	51.0	24151	51.1	CL ClinicalReports
1867	11.9	5590	11.8	CN Consultations
1120	7.2	3319	7.0	DS DischargeSummaries
335	2.1	878	1.8	ER EmergencyMedicine
2185	14.0	6789	14.4	HP HistoryAndPhysicals
1496	9.6	4534	9.6	OR OperativeReports
15617		47227		Total

Table 1: Distribution of Work-types

As features for categorization, we used a bag-of-words approach, but instead of the surface form of every token of a report, we used its semantic features as described in section 2.2. As a categorization engine, we used LIBSVM (Chang&Lin, 2001) with an RBF kernel. The features where weighted with TFIDF. In order to compensate for different document length, each feature vector was normalized to unit length. After some parameter tuning iterations, the SVM model performs really well with a microaveraged F1[1] value of 0.9437. This indicates high overall accuracy, and the macroaveraged F1 value of 0.9341 shows, that also lower frequency categories are predicted quite reliably. The detailed results are shown in table 2.

Thus the first step in the cascaded model, i.e. the selection of the work-type specific segment

[1] F1 $= \frac{2 \times precision \times recall}{precision + recall}$

	predicted								rec.	prec.	F1
true		CA	CL	CN	DS	ER	HP	OR			
CA	1966	1882	53	5	6	0	9	11	0.9573	0.9787	0.9679
CL	24151	25	23675	217	13	18	155	48	0.9803	0.9529	0.9664
CN	5590	1	447	4695	7	17	413	10	0.8399	0.8814	0.8601
DS	3319	1	37	8	3241	2	27	3	0.9765	0.9818	0.9792
ER	878	0	90	7	10	754	13	4	0.8588	0.9425	0.8987
HP	6789	4	512	393	22	7	5838	13	0.8599	0.9040	0.8814
OR	4534	10	31	2	2	2	3	4484	0.9890	0.9805	0.9847
										microaveraged	**0.9437**
										macroaveraged	**0.9341**

Table 2: Work-Type categorization results

topic model, yields reliable performance.

3.2 Segment Topic Classification

In contrast to work-type categorization, where whole reports need to be categorized, the identification of segment topics requires a different setup. Since not only the topic labels are to be determined, but also segment boundaries are unknown in the classification task, *each token* constitutes an example under this setting. Segments are then contiguous text regions with the same topic label. It is clearly not enough to consider only features of the token to be classified, thus we include also contextual and positional features.

3.2.1 Feature and Kernel Selection

In particular, we employ a sliding window approach, i.e. for each data set not only the token to be classified, but also the 10 preceding and the 10 following tokens are considered (at the beginning or towards the end of a report, context is reduced appropriately). This window defines the text fragment to be used for classifying the center token, and features are collected from this window again as described in section 2.2. Additionaly, the relative position (ranging from -1 to +1) of the center token is used as a feature.

The rationale behind this setup is that

1. usually topics in medical reports follow an ordering, thus relative position may help.
2. holding features also from adjacent segments might also be helpful since topic succession also follows typical patterns.
3. a sufficiently sized context might also smooth label assignment and prevent label oscillation, since the classification features for adjacent words overlap to a great deal.

A second choice to be made was the selection of the kernel best suited for this particular classification problem. In order to get an impression, we made a preliminary mini-experiment with just 5 reports each for training (4341 datasets) and testing (3382 datasets), the results of which are reported in table 3.

	Accuracy	
Feature Weight	linear	RBF
TFIDF	0.4977	0.3131
TFIDF normalized	0.5544	0.6199
Binary	0.6417	0.6562

Table 3: Preliminary Kernel Comparison

While these results are of course not significant, two things could be learned from the preliminary experiment:

1. linear kernels may have similar or even better performance,
2. training times with LIBSVM with a large number of examples may soon get infeasible (we were not able to repeat this experiment with 50 reports due to excessive runtime).

Since LibSVM solves linear and nonlinear SVMs in the same way, LibSVM is not particularly efficient for linear SVMs. Therefore we decided to switch to Liblinear (Fan et al., 2008), a linear classifier optimized for handling data with millions of instances and features[2].

[2] Indeed, training a model from 669 reports (463994 examples) could be done in less then 5 minutes!

				predicted class label (#)					
#	True Label	Total	F1	…	3	4	…	14	…
	…	…	…	…	…	…	…	…	…
3	Diagnosis	40871	0.603	…	24391	2864	…	8691	…
4	DiagAndPlan	21762	0.365	…	5479	6477	…	7950	…
	…	…	…	…	…	…	…	…	…
14	Plan	31729	0.598	…	5714	3419	…	21034	…
	…	…	…	…	…	…	…	…	…

Table 4: Confusion matrix (part of)

3.2.2 Segment Topic Classification Results

Experiments were performed on a randomly selected subset of reports from the "Consultations" work-type (1338) that were available both in corrected form and in raw ASR output form. Annotations were constructed for the corrected transcripts, as described in section 2.3, transfer of labels to the ASR output was performed as shown in section 2.3.2.

Both data sets were split into training and test sets of equal size (669 reports each), experiments with different feature weighting schemes have been performed on both corrected data and ASR output. The overall results are shown in table 5.

	corrected reports		ASR output	
Feature weights	micro-avg.F1	macro-avg.F1	micro-avg.F1	macro-avg.F1
TFIDF	0.7553	**0.5178**	0.7136	**0.4440**
TFIDF norm.	0.7632	0.3470	0.7268	0.3131
Binary	**0.7693**	0.4636	**0.7413**	0.3953

Table 5: Segment topic classification results

Consistently, macroaveraged F1 values are much lower than their microaveraged counterparts indicating that low-frequency topic labels are predicted with less accuracy.

Also, segment classification works better with corrected reports than with raw ASR output. The reason for that behaviour is

1. ASR data are more noisy due to recognition errors, and
2. while in corrected reports appropriate section headers are available (not as header, but the words) this is not necessarily the case in ASR output (also the wording of dictated headers and written headers may be different).

A general note on the used topic labels must also be made: Due to the nature of our data it was inevitable to use topic labels that overlap in some cases. The most prominent example here is "*Diagnosis*", "*Plan*", and "*Diagnosis and Plan*". The third label clearly subsumes the other two, but in the data available the physicians often decided to dictate diagnoses and the respective treatment in an alternating way, associating each diagnosis with the appropriate plan. This made it necessary to include all three labels, with obvious effects that could easily seen when inspecting the confusion matrix, a part of which is shown in table 4.

When looking at the misclassifications in these 3 categories it can easily be seen, that they are predominantly due to overlapping categories. Given these difficulties due to the data, the results are encouraging. There is, however, still plenty of room left for improvement.

3.3 Improving Topic Classification

Liblinear does not only provide class label predictions, it is also possible to obtain class probabilities. The usual way then to predict the label is to choose the one with the highest probability. When analysing the errors made by the segment topic classification task described above, it turned out that often the correct label was ranked second or third (cf. table 6). Thus, the idea of just taking the highest ranked class label could be possibly improved by a more informed choice.

While the segment topic classifier already takes contextual features into account, it has still no information on the classification results of the neighboring text segments. However, there are constraints on the length of text segments, thus, e.g. a text segment of length 1 with a different topic label than the surrounding text is highly implausible. Furthermore, there are also regularities in the succession of topic labels, which can be captured by the monostratal local classification only indirectly – if at all.

A look at table 7 exemplifies how a bet-

Label	count	correct prediction in		
		best	best 2	best 3
Allergies	3456	29.72	71.64	85.21
ChiefComplai	697			
Course	30			
Diagnosis	43565	64.69	83.29	91.37
DiagAndPlan	19409	35.24	70.45	86.81
DiagnosticSt	35554	82.47	91.34	93.05
Findings	791		0.38	1.26
Habits	2735	7.31	32.69	41.76
HistoryOfPre	122735	92.26	97.55	98.20
Medication	14553	85.87	93.38	95.22
Neurologic	5226	54.08	86.93	89.19
PastHistory	43775	71.13	86.26	88.82
PastSurgical	5752	49.32	78.88	84.47
PhysicalExam	86031	93.56	97.01	97.57
Plan	36476	62.57	84.63	94.65
Practitioner	1262	55.07	76.78	82.73
Procedures	109			
ReasonForEnc	15819	25.42	42.35	43.47
ReviewOfSyst	29316	79.81	89.90	91.87
Time	58			
Total	467349	76.93	88.65	92.00

Table 6: Ranked predictions

ter informed choice of the label could result in higher prediction accuracy. The segment labelled "*PastHistory*" correctly ends 4 tokens earlier than predicted, and, additionally, this label erroneously is predicted again for the phrase "*progressive weight loss*". The correct label, however, has still a rather high probability in the predicted label distribution. By means of stacking an additional classier onto the first one we hope to be able to correct some of the locally made errors a posteriori.

				Label probabilities (%)						
	True Label		Predicted	...	10	11	12	...	17	18
	...									
=	PastHistory [11]	age	PastHistory		0	95	0		0	0
=	PastHistory [11]	63	PastHistory		0	95	0		0	0
=	PastHistory [11]	and	PastHistory		0	95	0		0	1
=	PastHistory [11]	his	PastHistory		0	95	0		0	1
=	PastHistory [11]	father	PastHistory		0	88	0		0	9
=	PastHistory [11]	died	PastHistory		0	90	0		0	8
=	PastHistory [11]	from	PastHistory		0	84	0		0	14
=	PastHistory [11]	myocardial infa	PastHistory		0	81	0		0	17
=	PastHistory [11]	at	PastHistory		0	77	0		0	20
=	PastHistory [11]	age	PastHistory		0	78	0		1	19
=	PastHistory [11]	57	PastHistory		0	78	0		1	19
=	PastHistory [11]	period	PastHistory		0	78	0		1	19
-	ReviewOfSyst[18]	review	PastHistory		0	76	0		1	20
-	ReviewOfSyst[18]	of	PastHistory		0	76	0		1	21
-	ReviewOfSyst[18]	systems	PastHistory		0	78	0		0	19
-	ReviewOfSyst[18]	he	PastHistory		1	57	0		1	37
=	ReviewOfSyst[18]	has	ReviewOfSyst		1	32	0		1	58
=	ReviewOfSyst[18]	had	ReviewOfSyst		1	32	0		1	58
-	ReviewOfSyst[18]	progressive	PastHistory		1	49	0		1	42
-	ReviewOfSyst[18]	weight loss	PastHistory		1	60	0		1	32
=	ReviewOfSyst[18]	period	ReviewOfSyst		1	31	0		0	62
=	ReviewOfSyst[18]	his	ReviewOfSyst		1	13	0		1	81
=	ReviewOfSyst[18]	appetite	ReviewOfSyst		1	13	0		1	81
	...									

Table 7: predicted label probabilites

The setup for the error correction classifier we experimented with was as follows (it was performed only for the segment topic classifier trained on ASR output with binary feature weights):

1. The *training* set of the classifier was classified, and the predicted label probabilities were collected as features.

2. Again, a sliding window (with different sizes) was used for feature construction. Features were set up for each label at each window position and the respective predicted label probability was used as its value.

3. A linear classifier was trained on these features of the training set

4. This classifier was applied to the results of classifying the test set with the original segment topic classifier.

Three different window sizes were used on the corrected reports, only one window was applied on ASR output (cf. table 8). As can be seen, each context variant improved on both microaveraged and macroaveraged F1 in a range of 0,9 to 1.4 percent points. Thus, stacked error correction indeed is possible and able to improve classification results.

	corrected reports		ASR output	
context window	micro-avg.F1	macro-avg.F1	micro-avg.F1	macro-avg.F1
No correction	0.7693	0.4636	0.7413	0.3953
$[-3, +3]$	0.7782	0.4773	-	-
$[-6, +0]$	0.7798	0.4754	-	-
$[-3, +4]$	0.7788	0.4769	0.7520	0.4055

Table 8: A posteriori correction results

4 Conclusion and Outlook

We have presented a 3 step approach to segment topic identification in dictations of medical reports. In the first step, a categorization of work-type is performed on the whole report using SVM classification employing semantic features. The categorization model yields good performance (over 94% accuracy) and is a prerequisite for subsequent application of work-type specific segment classification models.

For segment topic detection, every word was assigned a class label based on contextual features in a sliding window approach. Here also semantic features were used as a means for feature generalisation. In various experiments, linear models using binary feature weights had the best performance. A posteriori error correction via classifier stacking additionally improved the results.

When comparing our results to the results of Jancsary et al. (2008), who pursue a multi-level segmentation aproach using conditional random fields optimizing over the whole report, the locally obtained SVM results cannot compete fully. On label chain 2, which is equivalent to segment topics as investigated here, Jancsary et al. (2008) report an estimated accuracy of 81.45 ± 2.14 % on ASR output (after some postprocessing), whereas our results, even with a posteriori error correction, are at least 4 percent points behind. This is probably due to the fact that the multi-level annotation employed in Jancsary et al. (2008) contains additional information useful for the learning task, and constraints between the levels improve segmentation behavior at the segment boundaries. Nevertheless, our approach has the merit of employing a framework that can be trained in a fraction of the time needed for CRF training, and classification works locally.

An investigation on how to combine these two complementary approaches is planned for the future. The idea here is to use the probability distributions on labels returned by our approach as (additional) features in the CRF model. It might be possible to leave out some other features currently employed in return, thereby reducing model complexity. The benefit we hope to get by doing so are shorter training time for CRF training, and, since, contrary to CRFs, SVMs are a large margin classification method, hopefully the CRF model can be improved by the present approach.

Acknowledgments

The work presented here has been carried out in the context of the Austrian KNet competence network COAST. We gratefully acknowledge funding by the Austrian Federal Ministry of Economics and Labour, and ZIT Zentrum fuer Innovation und Technologie, Vienna. The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Transport, Innovation, and Technology and by the Austrian Federal Ministry for Science and Research.

References

ASTM International. 2002. ASTM E2184-02: Standard specification for healthcare document formats.

C.-C. Chang and C.-J. Lin. 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9(2008):1871–1874.

C. Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.

M. Huber, J. Jancsary, A. Klein, J. Matiasek, H. Trost. 2006. Mismatch interpretation by semantics-driven alignment. *Proceedings of Konvens 2006*.

J. Jancsary, A. Klein, J. Matiasek, H. Trost. 2007. Semantics-based Automatic Literal Reconstruction Of Dictations. In Alcantara M. and Declerck T.(eds.), *Semantic Representation of Spoken Language 2007 (SRSL7)* Universidad de Salamanca, Spain, pp. 67-74.

J. Jancsary, J. Matiasek, H. Trost. 2008. Revealing the Structure of Medical Dictations with Conditional Random Fields. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 1–10.

T. Joachims. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference on Machine Learning. Springer, pp. 137–142.

D.A.B. Lindberg, B.L. Humphreys, A.T. McCray. 1993. The Unified Medical Language System. *Methods of Information in Medicine*, (32):281-291.

Lawrence Philips. 1990. Hanging on the metaphone. *Computer Language*, 7(12).

V.N. Vapnik 1995. *The Nature of Statistical Learning Theory*. Springer.