Detecting mentions of pain and acute confusion in Finnish clinical text

Hans Moen [1,2]*, Kai Hakala[1,3]*, Farrokh Mehryary[1,3], Laura-Maria Peltonen[2,4], Tapio Salakoski[1,5], Filip Ginter[1], Sanna Salanterä[2,4]

1. Turku NLP Group, Department of Future Technologies, University of Turku, Finland.
2. Department of Nursing Science, University of Turku, Finland.
3. The University of Turku Graduate School (UTUGS), University of Turku, Finland.
4. Turku University Hospital, Finland.
5. Turku Centre for Computer Science (TUCS), Finland.

firstname.lastname@utu.fi

Abstract

We study and compare two different approaches to the task of automatic assignment of predefined classes to clinical free-text narratives. In the first approach this is treated as a traditional mention-level named-entity recognition task, while the second approach treats it as a sentence-level multi-label classification task. Performance comparison across these two approaches is conducted in the form of sentence-level evaluation and state-of-the-art methods for both approaches are evaluated. The experiments are done on two data sets consisting of Finnish clinical text, manually annotated with respect to the topics pain and acute confusion. Our results suggest that the mention-level named-entity recognition approach outperforms sentence-level classification overall, but the latter approach still manages to achieve the best prediction scores on several annotation classes.

1 Introduction

In relation to patient care in hospitals, clinicians document the administrated care on a regular basis. The documented information is stored as clinical notes in electronic health record (EHR) systems. In many countries and hospital districts, a substantial portion of the information that clinicians document concerning patient status, performed interventions, thoughts, uncertainties and plans are written in a narrative manner using (natural) free text. This means that much of the patient information is only found in free-text form, as opposed to structured or coded information (c.f. standardized terminology, medications and diagnosis codes).

When it comes to information retrieval, management and secondary use, having the computer automatically identify and extract information from health records related to a given query or topic is desirable. This could, for example, be information about pain treatment given to a patient, or a patient group. Although free text is easy to produce by humans and allows for great flexibility and expressibility, it is challenging to have computers automatically classify and extract information from such text. The use of computers to automatically extract, label and structure information in free text is referred to as information extraction (Meystre et al., 2008), with named-entity recognition as a sub-task (Patawar Maithilee, 2015; Quimbaya et al., 2016). Due to the complexity of free text, this task is commonly approached using manually annotated text as training data for machine learning algorithms (see e.g. Velupillai and Kvist (2012)).

We present an ongoing work towards automated annotation of text, i.e. labelling with pre-defined classes/entity types, by first having the computer learn from a set of manually annotated clinical notes. The annotations concern two topics relevant to clinical care: *Pain* and *Acute Confusion*. To get a better insight into these topics and how this is being documented, two separate data sets have been manually annotated, one for each topic. For each of the two topics, a set of classes has been initially identified that reflect the information which the domain experts are interested in. An example sentence demonstrating the annotations is presented in Figure 1. The ultimate aim of this annotation work is to achieve improved documentation, assessment, handling and treatment of pain and acute confusion in hospitals (Heikkilä et al., 2016; Voyer et al., 2008). Now we want to inves-

*These authors contributed equally.

tigate how to best train the computer to automatically detect and annotate mentions of these topics in new, unseen text by exploring various machine learning methods.

We address this by testing and comparing two different overall approaches:

- Named-entity recognition (NER), where we have the computer attempt to detect the mention-level annotation boundaries.
- Sentence classification (SC), where we have the computer attempt to label sentences based on the contained annotations.

The motivation for comparing these two approaches is that: (a) the experts are satisfied with having the computer identify and extract information on sentence level; and (b) we hypothesize that several classes, in particular those reflecting the more complex concepts, are easier for the computer to identify when approached as a sentence classification task. Further, we are not aware of any other work where a similar comparison has been reported. The methods and algorithms that we explore are based on state-of-the-art machine learning methods for NER and SC.

2 Data

Pain is something that most patients experience to various degrees during or related to a hospital stay. Pain experience is subjective and hence it can be challenging for clinicians to properly assess if, how and to what extent patients are experiencing pain. Acute confusion is a mental state that patients may enter as a result of serious illness, infections, intense pain, anesthesia, surgery and/or drug use. When clearly evident, this is commonly diagnosed as acute confusion or delirium (Fearing and Inouye, 2009), which is identified as a mental disorder that affects perception, cognitivity, memory, personality, mood, psychomotricity and the sleep-wake rhythm. However, it can be challenging to clearly identify acute confusion or delirium at the point of care, in particular the milder cases. Still, signs and symptoms can often be found in the free text that clinicians document (Voyer et al., 2008), and the same goes for pain (Gunningberg and Idvall, 2007).

Our annotated data consists of a random sample of 280 care episodes that were gathered from patients who had an open heart surgery and who were admitted to one university hospital in Finland during the years 2005-2009. This sample includes 1327 days of nursing narratives and 2156 notes written by physicians. The same sample was used as data sets for both topics (i.e. pain and acute confusion). An ethical approval and an organizational permission from the hospital district was obtained before the data collection.

Separate annotation schemes, reflecting the classes and guidelines for the annotation work, were iteratively developed based on the literature for both topics. For pain the annotation scheme has 15 classes while the acute confusion scheme has 37 classes (see supplementary materials for more details). The annotation schemes were initially tested and refined by having the annotators annotate a separate data set of another 100 care episodes (not included in this study). The annotation task was conducted by four persons working in pairs of two, so that all the text was annotated by (at least) two annotators. This team of annotators consisted of two domain experts and two non domain experts with an informatics background. At the end, the annotators analyzed the made annotations with respect to common consensus before producing the final annotated data sets used in this study. The annotations were conducted using the brat annotation tool (Stenetorp et al., 2012).

The two data sets were individually divided into *training* (60%), *development* (20%) and *test* (20%) sets. As preprocessing of the data we tokenize and enrich the text with linguistic information in the form of lemmas and part-of-speech (POS) tags for each token. For this we use the Finnish dependency parser (Haverinen et al., 2014).

For training of word embeddings (word-level semantic vectors), we used a large corpus consisting of both physician and nursing narratives, extracted from the same university hospital (in Finland). In total, this corpus consist of approximately 0.5M nursing narratives and 0.4M physician notes, which amounts to 136M tokens.

3 Experiment and Methods

Below (Section 3.1 and 3.2) we describe the methods, algorithm implementations and hyper parameters used in the two approaches, i.e., named-entity recognition (NER) and sentence classification (SC). In the Results section, Section 4, we compare the scores achieved by these two ap-

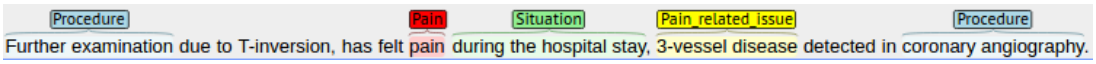

Figure 1: An artificial English example of the used pain annotations.

proaches for each of the two topics (i.e. pain and acute confusion).

3.1 Named-entity recognition (NER)

In this approach we focus on methods for predicting word-level annotation spans. More precisely we explore two such methods that have shown state-of-the-art performance in NER.

NERsuite Conditional random fields (CRFs) are a class of sequence modeling methods that have shown state-of-the-art performance in learning to identify biomedical named entities in text (Campos et al., 2013). We use a named-entity recognition toolkit called NERsuite (Cho et al., 2010), which is built on top of CRFsuite (Okazaki, 2007). For each of the two topics, one NERsuite model is trained using the corresponding training sets and the mentions are labeled using the common IOB tagging scheme. As training features, we use the original tokens, lemmas and POS tags. Although NERsuite allows the user to adjust regularization and label weight parameters, for this initial study we have used the default hyperparameters. It is worth noting that adjusting the regularization parameter is not as crucial for CRFs as it is for instance for support vector machines and strong results can be achieved even with the default values.

Several of the annotated entities have overlapping spans, e.g. the Finnish compound word *rintakipu* (chest pain) includes both *pain* and *location* mentions, but the standard CRF implementations are not able to do multi-label classification. Thus we form combination classes from the full spans of overlapping entities. This slightly distorts the annotated spans as the original mentions may have had only partial overlaps. Another option would have been to train separate models for each class, but as the number of classes is relatively high for both topics, this would have been very impractical.

CNN-BiLSTM-CRF The second method that we explore is an end-to-end neural model following the approach by Ma and Hovy (2016), which has produced state-of-the-art results for general domain English NER tasks. This model uses a CRF layer for the final predictions, but instead of relying on handcrafted features it utilizes a bidirectional recurrent neural network layer, with a long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997; Gers et al., 2000) chain, over input word embeddings. In addition to the input word embeddings, a convolutional layer is used over character embedding sequences to form another encoding for each token. Thus, this model is often called CNN-BiLSTM-CRF network. For training the model we use the example implementation provided by the authors [1].

Training the CNN-BiLSTM-CRF is computationally much more demanding then a standard CRF classifier and we have thus not ran an exhaustive hyperparameter search. Instead, we use the default values from the original paper except for setting the LSTM state dimensionality to 100 and learning rate to 0.05 as these produced slightly better results than the default values. The word embeddings are initialized with a word2vec (Mikolov et al., 2013) model trained on the large clinical Finnish text corpus.

3.2 Sentence classification (SC)

In this approach, we regard the task as a multi-label text classification task in which a sentence can be associated with multiple labels. For this task, we rely on artificial neural networks (ANN) since they have been shown to achieve state-of-the-art performance in text classification tasks (see e.g. Zhang et al. (2015); Tang et al. (2015)).

Neural network architecture We tried several neural network architectures, but report only the architecture that performed best. For both of the two topics, we apply a deep learning-based neural network architecture that use three separate LSTM chains: for the sequence of words, lemmas and POS tags.

The network has three separate channels for the words, lemmas and POS tags in the sentence. Each channel receives a sequence (words, lemmas or POS tags) as input. The items in the sequence are then mapped into their corresponding vector representations using a dedicated embedding look-up layer. The sequence of vectors is then input to an

[1] https://github.com/XuezheMax/LasagneNLP

LSTM chain and the last step-wise output of the chain is regarded as the representation of the sentence based on its words (or lemmas or POS tags).

Next, the outputs of the three channels are concatenated and the resulting vector is forwarded into the classification (decision) layer, which has a dimensionality equal to the number of annotation classes. The *sigmoid* activation function is applied on the output of the decision layer.

Training and optimization For implementation we use the Keras deep learning library (Chollet, 2015), with Theano tensor manipulation library (Bastien et al., 2012) as the back-end engine. We use *binary cross-entropy* as the objective function and the *Adam* optimization algorithm (Kingma and Ba, 2014) for training the network. We initialize the embeddings for words and lemmas with pre-trained vectors, trained using word2vec on the Finnish clinical corpus. For hyper-parameter optimization, we do a grid search and evaluate each model on the development set. To detect the best number of epochs needed for training, we use the *early stopping* method. Optimization is done against the *micro-averaged* F-score.

To avoid overfitting, we apply *dropout* (Srivastava et al., 2014) regularization with a rate of 20% on the input gates and with a rate of 1% on the recurrent connections of all LSTM units. In addition, we have set the dimensionality of the word, lemma and POS tag embeddings to 300 and the dimensionality of the LSTMs' output are also set to 300.

4 Results

We first evaluate the two NER methods on mention level using a strict offset matching criteria. The micro-averaged results are presented in Table 1. The NERsuite model achieves F-scores of 73.10 and 48.11 on the test sets of pain and acute confusion data set, respectively. Surprisingly the CNN-BiLSTM-CRF model is not able to reach the performance of the vanilla NERsuite on the pain dataset even though it is able to utilize pre-trained word embeddings. This might be due to the data sets being limited to open heart surgery patients and thus to a rather narrow vocabulary. Consequently we do not train CNN-BiLSTM-CRF on the confusion data. To analyse the performance of the NER approach in relation to the SC approach, we also convert the detected entity mentions to sentence-level predictions. For this the predictions from the best performing method, i.e. NERsuite, is used.

Approach	Precision	Recall	F-score
		Pain	
NERsuite	87.29	62.88	73.10
CNN-BiLSTM-CRF	79.30	63.80	70.71
		Acute confusion	
NERsuite	69.33	36.84	48.11

Table 1: Mention-level evaluation of the tested NER approaches on the test sets of the Pain and Acute confusion corpora. The reported numbers are micro-averaged over the various classes.

Table 2 shows the sentence-level scores for both the NER and SC approach. The best performing neural network used in the SC approach achieves slightly inferior results compared to the NER approach (when evaluated on sentence level). This seems to somewhat falsify our hypothesis about sentence-level classification methods potentially performing better than mention-level NER methods when the task is approached as a sentence classification task. Still, in Table 3 we see that the SC approach achieves best overall prediction scores for several of the annotation classes (see also supplementary materials). Based on our analysis so far, it is difficult to say whether these classes (i.e. the concepts they represents) are more "complex" than the others, or if there are some other factors affecting the results. In an attempt to achieve better insight into this, we calculated the average annotation spans and vocabulary size associated with the different classes. However, these numbers did not show any clear trend.

Approach	**Pain**	**Acute confusion**
NER	78.61	59.41
SC	77.65	57.49

Table 2: Micro-averaged F-scores for the different approaches on the test sets of the pain and acute confusion data sets. NERsuite was used to produce the NER scores.

The actual pain mentions which are divided into explicit, implicit and potential pain subcategories all achieve relatively high performance, implicit pain being the hardest to predict (see supplementary materials for more details). The other classes, which describe additional information about the pain mentions, are generally speaking harder to detect than the actual pain mentions. The acute confusion related entities seems to be much harder

Approach	Pain	Acute confusion
NER	8	11
SC	7	8
Equal performance	0	18

Table 3: Counts showing the number of classes that the various approaches performed best at predicting.

to predict due to the vague and sparse nature of these concepts.

5 Discussion and Future Work

In this study we have gathered the initial results for detecting mentions of pain and acute confusion in Finnish clinical text. We also use a relaxed evaluation based on sentence level predictions and experiment with approaches designed specifically for this definition. Surprisingly the NERsuite based mention-level approach outperforms all other tested methods, showing strong performance and being the best suited alternative for real-world applications. However, it might be that these two approaches are complementary.

As the used datasets are limited to open heart surgery patients, a critical future work direction will be assessing the generalizability of the trained models on larger sets of patient health records, and from other hospital units. This study also reveals that multiple classes in the annotation schemes, in particular for acute confusion, need more manual annotation data, i.e. more training examples, in order to be reliably detected in an automatic manner.

As many of the classes can be considered as descriptive attributes of the pain and acute confusion mentions, but the relations have not been annotated explicitly, another future work direction is to investigate how often these relations are ambiguous and whether the relation extraction could be solved in an unsupervised fashion.

Acknowledgments

Funding sources for this research were Academy of Finland and Tekes (Räätäli ("Tailor") project). We would like to thank the annotators, Pauliina Anttila, Timo Viljanen, Satu Poikajärvi and Kristiina Heikkilä. We would also like to thank Juho Heimonen for assisting us in preprocessing the clinical text.

References

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. 2012. Theano: New features and speed improvements. *arXiv preprint arXiv:1211.5590 (2012)* .

David Campos, Sérgio Matos, and José Luís Oliveira. 2013. Gimli: open source and high-performance biomedical name recognition. *BMC Bioinformatics* 14(1):54.

HC Cho, N Okazaki, M Miwa, and J Tsujii. 2010. Nersuite: a named entity recognition toolkit. `http://nersuite.nlplab.org`. Last visited 20th April 2017.

Franois Chollet. 2015. Keras. `https://github.com/fchollet/keras`. Last visited 10th March 2017.

Michael A Fearing and Sharon K Inouye. 2009. Delirium. *Focus* 7(1):53–63. https://doi.org/10.1176/foc.7.1.foc53.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with LSTM. *Neural Computation* 12(10):2451–2471.

Lena Gunningberg and Ewa Idvall. 2007. The quality of postoperative pain management from the perspectives of patients, nurses and patient records. *Journal of Nursing Management* 15(7):756–766.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missil, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for finnish: the turku dependency treebank. *Language Resources and Evaluation* 48:493–531. https://doi.org/10.1007/s10579-013-9244-1.

Kristiina Heikkilä, Laura-Maria Peltonen, and Sanna Salanterä. 2016. Postoperative pain documentation in a hospital setting: A topical review. *Scandinavian Journal of Pain* 11:77–89.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://arxiv.org/abs/1412.6980.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. *arXiv preprint arXiv:1603.01354* .

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. 2008. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics* 35:128–44.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Naoaki Okazaki. 2007. Crfsuite: A fast implementation of Conditional Random Fields (CRFs). http://www.chokkan.org/software/crfsuite/. Last visited 20th April 2017.

Potey M. A. Patawar Maithilee. 2015. Approaches to named entity recognition: A survey. *International Journal of Innovative Research in Computer and Communication Engineering* 3(12):12201–12208.

Alexandra Pomares Quimbaya, Alejandro Sierra Mnera, Rafael Andrs Gonzlez Rivera, Julin Camilo Daza Rodrguez, Oscar Mauricio Muoz Velandia, Angel Alberto Garcia Pea, and Cyril Labb. 2016. Named entity recognition over electronic health records through a combined dictionary-based approach. *Procedia Computer Science* 100:55 – 61. https://doi.org/http://dx.doi.org/10.1016/j.procs.2016.09.123.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15:1929–1958. http://jmlr.org/papers/v15/srivastava14a.html.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 102–107.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 1422–1432.

Sumithra Velupillai and Maria Kvist. 2012. Fine-grained certainty level annotations used for coarser-grained E-Health scenarios. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, Springer Berlin Heidelberg, volume 7182 of *Lecture Notes in Computer Science*, pages 450–461. https://doi.org/10.1007/978-3-642-28601-8_38.

Philippe Voyer, Martin G Cole, Jane McCusker, Sylvie St-Jacques, and Johanne Laplante. 2008. Accuracy of nurse documentation of delirium symptoms in medical charts. *International Journal of Nursing Practice* 14(2):165–177. https://doi.org/10.1111/j.1440-172X.2008.00681.x.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc., pages 649–657.

A Supplemental Material

The supplementary material includes specific information about the annotation classes for the pain and acute confusion data sets, as well as the detailed evaluation of the studied methods.

English names	Finnish names	SC	NER
A recurrent situation	Toistuva_tilanne	83.60	86.10
Care plan	Suunnitelma	83.39	84.24
Implicit pain	Implisiittinen_kipu	74.64	73.68
Pain related issue	Kipuun_liittyva_asia	54.69	54.51
Location of pain	Sijainti	83.36	87.23
Pain	Kipu	92.04	93.23
Pain intensity	Voimakkuus	75.29	80.80
Pain management	Kivunhoito	85.36	86.97
Patient education	Ohjeistus	33.33	0.00
Potential pain	Potentiaalinen_kipu	93.69	95.58
Procedure	Toimenpide	64.73	63.90
Quality of pain	Laatu	59.33	71.07
Success of treatment	Hoidon_onnistuminen	73.74	64.56
Situation	Tilanne	37.21	28.57
Time	Aika	79.49	72.41
	Micro-average	**77.65**	**78.61**

Table 4: Comparison of SC and NER for sentence classification, for pain corpus test set, evaluated on micro-averaged F1-scores.

English Names	Finnish Names	SC	NER
Abnormal level of consciousness	Muu_poikkeava_tajunnan_taso	28.57	0.00
Aggressiveness	Aggressiivisuus_vihaisuus	20.00	64.00
Appetite disturbance	Ruokahalun_hairio	57.35	59.02
Calming activity	Rauhoittelu	0.00	0.00
Confusion	Sekavuus	85.95	95.00
Delirium	Delirium	0.00	0.00
Delusion	Harhaisuus	34.29	27.59
Dementia	Dementia	0.00	0.00
Desorientation	Desorientaatio	66.67	89.86
Diagnosed	Diagnosoitu	0.00	0.00
Disturbance in ability to focus	Vaikea_kiinnittaa_huomiota	15.39	0.00
Disturbance in the quality of speech	Puheen_laadun_hairiot	40.00	29.41
Drowsy	Unelias	77.98	79.44
Falls - fall out of bed	Kaatuminen_Sangysta_tippuminen	0.00	0.00
Hyper-alert	Ylivalpas	0.00	0.00
Hyperactivity	Hyperaktiivisuus	68.71	70.23
Hypoactivity	Hypoaktiivisuus	26.67	22.22
Infusion line detachment	Letkun_irtoaminen	20.00	20.00
Memory disorder	Muistiongelma	73.24	80.00
Not awakable	Ei_herateltavissa	0.00	0.00
Orientation to time and place	Orientoiminen_aikaan_paikkaan	0.00	0.00
Other abnormal behavior	Muu_poikkeava_kayttaytyminen	0.00	0.00
Other affective disturbance	Muu_tunnehairio	64.52	42.25
Other care activity	Muu_hoitotoimenpide	0.00	0.00
Other cognitive disturbance	Muu_kognitiivinen_hairio	0.00	0.00
Other disturbance of attention	Muu_tarkkaavaisuuden_hairio	0.00	0.00
Other incident	Muu_haittatapahtuma	0.00	0.00
Other symptom	Muu_oire	0.00	0.00
Pain management	Kivunhoito	51.52	61.33
Problems with motor functions	Motoriikan_ongelmat	59.56	59.79
restraint - restraining	Lepositeet_sitominen	75.00	76.19
Sleep-wake disorder	Unirytmin_valverytmin_hairiot	54.32	48.65
Slow rate of speech - Speechlessness	Hidastunut_puhe_puhumattomuus	0.00	0.00
Substance induced delirium	Substance_induced_delirium	0.00	0.00
Unappropriate behaviour	Asiaankulumaton_kayttaytyminen	9.52	10.26
Uncertain	Epavarma	0.00	0.00
Unorganised thinking	Ajatuksenkulun_jarjestaytymattomyys	25.81	21.43
	Micro-average	**57.49**	**59.41**

Table 5: Comparison of SC and NER for sentence classification, for acute confusion corpus test set, evaluated on micro-averaged F1-scores.

English Names	Finnish Names	Train	Devel	Test	Total
A recurrent situation	Toistuva_tilanne	589	215	210	1014
Care plan	Suunnitelma	517	176	170	863
Implicit pain	Implisiittinen_kipu	552	160	201	913
Pain related issue	Kipuun_liittyva_asia	1058	377	372	1807
Location of pain	Sijainti	1001	326	333	1660
Pain	Kipu	1655	536	549	2740
Pain intensity	Voimakkuus	1094	291	341	1726
Pain management	Kivunhoito	1158	368	419	1945
Patient education	Ohjeistus	11	3	4	18
Potential pain	Potentiaalinen_kipu	752	222	255	1229
Procedure	Toimenpide	1423	478	468	2369
Quality of pain	Laatu	323	100	125	548
Success of treatment	Hoidon_onnistuminen	226	75	102	403
Situation	Tilanne	286	85	82	453
Time	Aika	1257	386	426	2069
	Overall	11902	3798	4057	19757
	Tokens	437935	147444	153975	739354
	Sentences	71390	23470	25123	119983
	Documents	2084	697	702	3483

Table 6: Pain annotation counts per class.

English Names	Finnish Names	Train	Devel	Test	Total
Abnormal level of consciousness	Muu_poikkeava_tajunnan_taso	11	9	6	26
Aggressiveness	Aggressiivisuus_vihaisuus	24	5	16	45
Appetite disturbance	Ruokahalun_hairio	229	84	76	389
Calming activity	Rauhoittelu	6	4	6	16
Confusion	Sekavuus	131	45	60	236
Delirium	Delirium	4	1	1	6
Delusion	Harhaisuus	37	16	25	78
Dementia	Dementia	3	2	1	6
Desorientation	Desorientaatio	77	25	38	140
Diagnosed	Diagnosoitu	1	0	0	1
Disturbance in ability to focus	Vaikea_kiinnittaa_huomiota	29	8	12	49
Disturbance in the quality of speech	Puheen_laadun_hairiot	43	10	25	78
Drowsy	Unelias	275	88	115	478
Falls - fall out of bed	Kaatuminen_Sangysta_tippuminen	6	0	3	9
Hyper-alert	Ylivalpas	3	1	1	5
Hyperactivity	Hyperaktiivisuus	232	66	78	376
Hypoactivity	Hypoaktiivisuus	103	35	44	182
Infusion line detachment	Letkun_irtoaminen	15	4	9	28
Memory disorder	Muistiongelma	92	40	41	173
Not awakable	Ei_herateltavissa	15	7	6	28
Orientation to time and place	Orientoiminen_aikaan_paikkaan	6	0	0	6
Other abnormal behavior	Muu_poikkeava_kayttaytyminen	6	2	4	12
Other affective disturbance	Muu_tunnehairio	109	52	52	213
Other care activity	Muu_hoitotoimenpide	12	9	7	28
Other cognitive disturbance	Muu_kognitiivinen_hairio	23	4	5	32
Other disturbance of attention	Muu_tarkkaavaisuuden_hairio	10	1	3	14
Other incident	Muu_haittatapahtuma	10	3	5	18
Other symptom	Muu_oire	25	5	8	38
Pain management	Kivunhoito	118	39	40	197
Problems with motor functions	Motoriikan_ongelmat	329	93	117	539
restraint - restraining	Lepositeet_sitominen	25	8	13	46
Sleep-wake disorder	Unirytmin_valverytmin_hairiot	147	56	48	251
Slow rate of speech - Speechlessness	Hidastunut_puhe_puhumattomuus	25	11	11	47
Substance induced delirium	Substance_induced_delirium	1	0	0	1
Unappropriate behaviour	Asiaankulumaton_kayttaytyminen	81	22	33	136
Uncertain	Epavarma	1	0	0	1
Unorganised thinking	Ajatuksenkulun_jarjestaytymattomyys	62	17	24	103
	Overall	2326	772	933	4031
	Tokens	434542	149387	155425	739354
	Sentences	71146	23797	25040	119983
	Documents	2080	698	705	3483

Table 7: Acute confusion annotation counts per class.