# TAKELAB: Medical Information Extraction and Linking with MINERAL

**Goran Glavaš**
University of Zagreb
Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
goran.glavas@fer.hr

## Abstract

Medical texts are filled with mentions of diseases, disorders, and other clinical conditions, with many different surface forms relating to the same condition. We describe *MINERAL*, a system for extraction and normalization of disease mentions in clinical text, with which we participated in the Task 14 of SemEval 2015 evaluation campaign. MINERAL relies on a conditional random fields-based model with a rich set of features for mention detection, and a semantic textual similarity measure for entity linking. MINERAL reaches joint extraction and linking performance of 75.9% relaxed $F_1$-score (strict score of 72.7%) and ranks fourth among 16 participating teams.

## 1 Introduction

Clinical narratives contain numerous mentions of diseases and disorders. Recognizing these mentions in text and normalizing the different superficial forms of a disorder to the same canonical form could enable new types of analyses that would be beneficial for both medical professionals and patients.

Detection and normalization of various concepts such as named entities (McCallum and Li, 2003; Krishnan and Manning, 2006) or events (Bethard, 2013; Glavaš and Šnajder, 2014) has long been in the focus of the NLP community. Disorder mentions in clinical text, however, have some peculiarities not typical for traditional information extraction tasks such as discontinuity or distributivity of a single token to multiple disorder mentions. For example, the snippet

> *"Patient's extremities were turned in and clinched together as a consequence of. . . "*

contains two mentions of medical conditions, *"extremities turned in"* and *"extremities clinched together"*, which share the token *"extremities"*, with the latter mention being discontinuous.

In this paper we present the *MINERAL* (Medical INformation ExtRAction and Linking) system for recognizing and normalizing mentions of clinical conditions, with which we participated in Task 14 of SemEval 2015 evaluation campaign. The system recognizes disorder mentions via the supervised conditional random fields (CRF) model with a rich set of lexical, gazetteer-based, and informativeness-based features. We apply a set of post-processing rules to construct disorder mentions from token-level annotations which follow the BEGIN-INSIDE-OUTSIDE scheme. We utilize a measure of semantic textual similarity to link recognized disorder mentions to entries in the SNOMED-CT medical database. Our approach is resource light in the sense that, except for SNOMED-CT which is necessary for normalization, it does not rely on medical NLP resources.

We ranked fourth (relaxed evaluation setting) among 16 teams in the official evaluation, with 3% lower performance than the best-performing system. Such a result suggests that coupling sequence labelling for mention recognition with an STS measure for concept normalization poses a viable solution for entity recognition in the clinical domain. We make the MINERAL system freely available.[1]

[1] http://takelab.fer.hr/mineral

*Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 389–393,
Denver, Colorado, June 4-5, 2015.

## 2 Clinical Information Extraction

Clinical concept extraction is an essential task in medical natural language processing. While early approaches heavily relied on domain-specific vocabularies (Friedman et al., 1994; Aronson, 2001; Zeng et al., 2006), more recent efforts leverage the human-annotated corpora to develop machine learning models for the extraction of medical concepts (Tang et al., 2013; Uzuner et al., 2010). The rise in the number of data-driven efforts in the medical domain was particularly motivated by the shared tasks such as i2b2 challenges (Uzuner et al., 2010) and ShARe/CLEF eHealth Evaluation Lab (Suominen et al., 2013).

The first subtask of the SemEval Task 14, in which we participated, was essentially the same as the first task in the ShARe/CLEF eHealth campaign. We did not participate in the second subtask on extracting arguments of disorder mentions. The best performing system of the ShARe/CLEF eHealth task on disorder extraction and normalization (Tang et al., 2013) employed CRF and structured SVM models for mention extraction and the traditional vector-space model from information retrieval (Salton et al., 1975) for disorder normalization.

Similar to (Tang et al., 2013), we employ the CRF model for extraction of disorder mentions, but we leverage recent findings in word vector representations (Mikolov et al., 2013) for feature computation. We make use of the state-of-the-art measure of semantic similarity of short texts (Šarić et al., 2012) for concept normalization.

## 3 MINERAL

MINERAL consists of two subsystems: one for extracting disorder mentions and the other for normalizing extracted mentions by assigning them a Concept Unique Identifier (CUI) from the SNOMED-CT database (Stearns et al., 2001).

### 3.1 Disorder Mention Extraction

At the core of the extraction subsystem is the CRF model with lexical, gazetteer-based, and informativeness-based features. We decided to use the BEGIN-INSIDE-OUTSIDE annotation scheme for the CRF model, although this scheme does not account for token-sharing disorder mentions. Thus, we apply a set of postprocessing rules to derive disorder mentions from token-level outputs produced by the CRF model and to handle most frequent cases of token-sharing mentions (e.g., *"abdomen non-disturbed and non-distended"*).

#### 3.1.1 Features

We feed the CRF model with a rich set of features that can be divided into (1) token-based features, (2) gazetteer-based features, and (3) information content-based features. All of the features are templated on the symmetric window of size two, i.e., computed for two preceding tokens, current token, and two subsequent tokens.

**Token-based features (TK).** Token-based features group all features which can be computed just from the token at hand. These include the surface form, lemma, stem, POS-tag, and shape (encoding of the capitalization of the word, e.g., "UL" for "Atrial") of the word. We also encode the first and the last character bigram and trigram of the word as features.

**Gazetter-based features (GZ).** Features in this group rely on comparison of tokens in text with entries in the SNOMED-CT database and with disease annoations on the training set. For each token we compute: the maximum similarity with any of the words (1) *starting* a SNOMED-CT entry, (2) *inside* a SNOMED-CT entry, and (3) *ending* a SNOMED-CT entry. We compute the same three features only considering gold annotations in the training set as gazetteer entries. We compute the semantic similarity between two words as the cosine between their corresponding word embedding vectors. We trained the embedding vectors with the word2vec tool (Mikolov et al., 2013) on the large unlabeled corpus of clinical texts (with over 400K documents) provided by the task organizers. We also counted the number of gazetteer entries that start with, contain, and end with the token at hand.

**Information content-based features (IC).** These features compute the informativeness of ngrams within the clinical domain and compare it their general informativeness. We use *information content* as a measure of the informativeness of the word $w$ within a corpus $C$:

$$ic(w) = -\log \frac{\mathit{freq}(w) + 1}{\sum_{w' \in C} \mathit{freq}(w') + 1}$$

where $freq(w)$ is the frequency of the word $w$ in corpus $C$. We compute three different information content-based features. First, we compute the information content of the word within a large corpus of clinical narratives. Secondly, we compute the ratio of the information content of the word computed on the clinical corpus and the information content of the same word computed on a large general corpus. We used Google Books ngrams (Michel et al., 2011) as the general corpus. The rationale here is that the clinical concepts such as diseases and disorders will have a higher relative frequency and, consequently, lower information content in the clinical corpus than in the general corpus. Finally, the third feature we compute is the mutual information of the bigrams in the clinical corpus, which we define via the information content:

$$mi(w_1, w_2) = \frac{ic(w_1 w_2)}{ic(w_1) \cdot ic(w_2)}$$

where $ic(w_1 w_2)$ is the information content of the bigram $w_1 w_2$. Mutual information score indicates pairs of words that often appear together (e.g., *"atrial dilatation"*). For each word $w_i$ we compute the mutual information of the bigrams it constitues with the previous word (i.e., $w_{i-1} w_i$) and the subsequent word (i.e., $w_i w_{i+1}$).

### 3.1.2 Postprocessing

The only reasonable postprocessing strategy with the B-I-O scheme is to join each INSIDE token with the closest preceding BEGIN token. However, this strategy requires rule-based fixes for common situations in which two disorder mentions share a token. We designed postprocessing rules by observing the most frequent mistakes our CRF model made on the development set provided by the organizers. This led to three particular fixes: (1) mentions of *abdomen condition* typically correspond to two disorder mentions sharing the token *"abdomen"* (e.g., processing *"abdomen non-tender and non-distended"* results with two disorder mentions – *"abdomen non-tender"* and *"abdomen non-distended"*); (2) mentions of *allergies* typically share the token *"allergies"* (e.g., processing *"Allergies: Roxicet / Penicillins / Aspirin"* produces three mentions – *"Allergies Roxicet"*, *"Allergies Penicillins"*, and *"Allergies Aspirin"*); and (3) the CRF model rather frequently fails to recognize the type of the *hepatitis*. We associate the type of the *hepatitis* (e.g., *"B"*) found in the proximity of the token *"hepatitis"* when CRF fails to do so.

## 3.2 Mention Normalization

The normalization subsystem assigns a CUI to each extracted disorder mention by comparing the semantic similarity of the mention with the SNOMED-CT entries. Given that SNOMED-CT has over 650K entries, it is infeasible to compute the similarity of the disorder mentions with all database entries. Therefore, we first filtered out only the entries which contain at least one lemma from the extracted mention. E.g., for the mention *"melena due to gastrointestinal haemorrhage"* we would consider only the SNOMED-CT entries containing either *"melena"*, *"gastrointestinal"*, or *"haemorrhage"*.

We compute the similarity as the modified variant of the *greedy weighted alignment overlap* (GWAO) measure from (Šarić et al., 2012). To compute this score, we iteratively pair the words – one from extracted mention and the other from the database entry – according to their semantic similarity. In each iteration we greedily select the pair of words with the largest semantic similarity, and remove these words from their corresponding text snippets. The similarity between words is computed as the cosine between their embedding vectors obtained with `word2vec` (Mikolov et al., 2013) on the large unlabeled corpus of clinical narratives. Let $P(m, s)$ be the set of word pairs obtained through the alignment between the extracted mention $m$ and the SNOMED-CT entry $s$ and let $vec(w)$ be the embedding vector of the word $w$. The GWAO score is then computed as follows:

$$gwao(m, s) = \sum_{\substack{(w_m, w_s) \\ \in P(m,s)}} \alpha \cdot \cos\left(vec(w_m), vec(w_s)\right)$$

where $\alpha$ is the larger of the information contents of the two words, $\alpha = \max\left(ic(w_m), ic(w_s)\right)$. The $gwao(m, s)$ score is normalized with the sum of information contents of words from $m$ and $s$, respectively, and the harmonic mean of the two normalized scores is the final similarity score. We assign to the extracted mention the CUI of the most similar SNOMED-CT entry, assuming the similarity is above some treshold $\lambda$ (otherwise, the label "CUI-less" is assigned to the mention). The optimal value of $\lambda$ is

|  | Strict |  |  | Relaxed |  |  |
|---|---|---|---|---|---|---|
| Model | P | R | $F_1$ | P | R | $F_1$ |
| *TK* | 75.6 | 65.6 | 70.2 | 90.0 | 80.4 | 84.9 |
| *TK + GZ* | 75.1 | 66.1 | 70.3 | 89.6 | 80.9 | 85.0 |
| *TK + IC* | 76.4 | 66.3 | 71.0 | 90.2 | 80.4 | 85.1 |
| *All feat.* | 76.3 | 66.9 | 71.3 | 90.1 | 81.1 | 85.4 |
| *All + PPR* | 77.4 | 69.1 | 73.0 | 90.1 | 82.2 | 86.0 |

Table 1: Model selection results.

|  | Strict |  |  | Relaxed |  |  |
|---|---|---|---|---|---|---|
| Team | P | R | $F_1$ | P | R | $F_1$ |
| ezDI | 78.3 | 73.2 | 75.7 | 81.5 | 76.1 | 78.7 |
| ULisboa | 77.9 | 70.5 | 74.0 | 80.6 | 72.9 | 76.5 |
| UTH-CCB | 77.8 | 69.6 | 73.5 | 79.7 | 71.4 | 75.3 |
| UWM | 77.3 | 69.9 | 73.4 | 80.9 | 73.1 | 76.8 |
| **TakeLab** | **76.1** | **69.6** | **72.7** | **79.4** | **72.7** | **75.9** |
| Bioinf.-UA | 69.0 | 73.6 | 71.2 | 71.9 | 76.6 | 74.2 |

Table 2: Official SemEval Task 14 (subtask 1) evaluation.

determined by maximizing the CUI prediction accuracy on the training and development set. A useful add-on to the normalization step is the memorization of CUIs for all disorder mentions observed in the training set. In other words, a memorized mention observed in the test set will be assigned the CUI it had in the training set.

## 4 Evaluation

Participants were provided with a training set consisting of 298 clinical documents and a development set with 133 documents. We used the training and development set to optimize the model (features, postprocessing rules, and the similarity treshold $\lambda$). A test set of 100 clinical documents was used for official evaluation.

### 4.1 Model Optimization

We trained the CRF model with different combinations of feature groups (TK, GZ, and IC) and evaluated the performance of these models on the development set. We also evaluated the contribution of the postprocessing rules (PPR) on the development set. The extraction performance of the different models is shown in Table 4.1. The model using only token-based features alone (model TK) achieves solid performance. Information content-based features (model TK + IC) seem to have a more positive impact on the performance than the gazetteer-based features (model TK + GZ). Still, the model with all features displays the best performance. Applying postprocessing rules further boosts the performance on the development set, which is expected, because the rules were designed precisely to fix the most frequent errors on that dataset. We submitted the model *All + PPR* for official evaluation. We also optimized the similarity treshold $\lambda$ to maximize the normalization accuracy on the development set, selecting the optimal value of $\lambda = 0.83$.

### 4.2 Official Results

A subset of the official ranking on the test set is shown in Table 4.2. MINERAL ranks fourth among 16 teams in relaxed evaluation and fifth in strict evaluation, with only 3% lower $F_1$ performance than the best performing system.

Like most other systems, MINERAL displays higher precision than recall. This would suggest a non-negligible amount of obdurate disorder mentions which appear rarely in clinical documents and which are not semantically similar with more frequent disorders.

## 5 Conclusion

We described MINERAL, a system for extraction and normalization of disorder mentions in clinical text, with which we participated in Task 14 of SemEval 2015. At the core of the mention extraction approach is the CRF model built on B-I-O annotation scheme and a rich set of lexical, gazetteer-based, and informativeness-based features. We link the disease mentions to the SNOMED-CT entries using a measure of semantic textual similarity of short texts.

MINERAL achieved performance of almost 76% $F_1$ (relaxed evaluation setting), ranking us fourth out of 16 teams participating in the task, with 3% lower performance than the best-performing team. Such a result suggests that a resource light approach with sequence labeling (with semantic features) for mention extraction and STS measures for concept normalization offers competitive performance in the clinical domain.

## References

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17.

Steven Bethard. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Second Joint Conference on Lexical and Computational Semantics (StarSEM)*, volume 2, pages 10–14.

Carol Friedman, Philip O. Alderson, John H.M. Austin, James J. Cimino, and Stephen B. Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174.

Goran Glavaš and Jan Šnajder. 2014. Construction and evaluation of event graphs. *Natural Language Engineering*, pages 1–46.

Vijay Krishnan and Christopher D. Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 1121–1128.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 188–191.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, and Jon Orwant. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

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*, pages 3111–3119.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. TakeLab: Systems for measuring semantic text similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval)*, pages 441–448.

Michael Q. Stearns, Colin Price, Kent A. Spackman, and Amy Y. Wang. 2001. SNOMED Clinical Terms: Overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W. Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R. South, Danielle L. Mowery, and Gareth J.F. Jones. 2013. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In *Information Access Evaluation: Multilinguality, Multimodality, and Visualization*, pages 212–231.

Buzhou Tang, Yonghui Wu, Min Jiang, Joshua C. Denny, and Hua Xu. 2013. Recognizing and encoding discorder concepts in clinical text using machine learning and vector space model. In *Workshop of ShARe/CLEF eHealth Evaluation Lab 2013*.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Qing T. Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N. Murphy, and Ross Lazarus. 2006. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making*, 6(1):30.