Clinical Data Classification using Conditional Random Fields and Neural Parsing for Morphologically Rich Languages

Razieh Ehsani, Tyko Niemi, Gaurav Khullar and Tiina Leivo
Digital Workforce Services Oy, Helsinki, Finland
{name.lastname}@digitalworkforce.fi

Abstract

Past prescriptions constitute a central element in patient records. These are often written in an unstructured and brief form. Extracting information from such prescriptions enables the development of automated processes in the medical data mining field. This paper presents a Conditional Random Fields (CRFs) based approach to extract relevant information from prescriptions. We focus on Finnish language prescriptions and make use of Finnish language specific features. Our labeling accuracy is 95%, which compares favorably to the current state-of-the-art in English language prescriptions. This, to the best of our knowledge, is the first such work for the Finnish language.

1 Introduction

Processing and mining unstructured data is a major contemporary challenge. Automated methods reduce human labor and increase accuracy and proficiency. Application of such methods revolutionized many processes in the healthcare sector by eliminating huge amounts of manual work needed to process archive files. Automated processing of past patient data, such as prescriptions, allows easy digital access to patient records and allows healthcare practitioners to quickly inquire about family history, past medication usage, and other important data.

A large number of medical archives are in text format. Prescriptions, clinical reports, and other clinical texts are widely available but the problem with most of these texts is that they are unstructured and cannot be processed into a structured database directly. Extracting information from these is an important data mining problem called clinical text analysis.

In this paper, we will introduce an approach to extract entities from prescriptions. These entities are dosage, dosage unit and frequency. All prescriptions are in the Finnish language. Finnish is an agglutinative language with rich derivational and inflectional morphology. Morphemes mostly come afterword stem as suffixes and phonetics may also change depending on the morphemes. Finnish has complex vowel harmony and consonant gradation processes which causes large variations in each word stem.

This paper is organized as follows: In Section 2 we briefly introduce some important works as related works. Section 3 is about data that we used for training and also is about prepossessing step. In Section 4 we give information about the model and approach that we used in the paper. In Section 5 we present experimental results and we discuss over different tests. In Section 6 we describe post-processing step for mapping extracted information from prescriptions to the standardized master table data. Finally in Section 7 we conclude this paper.

2 Related work

CRFs are widely used in agglutinative language processing and have good accuracy when linguistic features are used (Yıldız et al., 2015), (Ehsani et al., 2012).

Here we list some of the existing tools in medication extraction. MedLEE (Friedman, 2000) is one which uses handwritten rules for extracting and encoding and structuring clinical information using free-form texts like patient reports. MetaMap (Aronson and Lang, 2010) also is a rule-based tool which extracts medication names by querying in the Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004).

Patrik et.al. (Patrick and Li, 2009) uses CRFs and also rule based approach to extract information form i2b2 data (Uzuner et al., 2011). Halgrim et.al. (Li et al., 2010) uses CRFs with simple fea-

Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 149–155
Minneapolis, Minnesota, June 7, 2019.

tures like n-grams and length of words over a small dataset from i2b2 task. They use a rule-based algorithm to improve the accuracy of CRF classification.

Another work (Tao et al., 2017) is also related to the same i2b2 shared task. They use CRFs to extract dosage unit, dosage and frequency. They show that CRFs performs better than other classifiers. They also try adding word embedding to their model but there is no significant improvement in dosage, dosage unit and frequency labeling. They employ POS tags of tokens besides some categorical features.

3 Data and preprocessing

Our training data contains 9692 prescriptions. We annotate these prescriptions to 4 categories: i) Dosage, which shows the amount of dosage of medication ii) Dosage unit gives the unit of medication, like "tablet" iii) Frequency of using dosage can be more than one token and iv) Comments category is for all other tokens in prescription. We annotated data manually by Finnish native speakers and with the supervision of healthcare professionals.

As mentioned before, working with the Finnish language brings its own challenges, we now discuss these in more detail. Beside compound words, rich morphology and phonology of Finnish language means that the same root word can appear in vastly different forms in texts. In addition to that, the colloquial patient-friendly language of the prescriptions means that they don't perfectly follow grammatical rules or spelling.

For example, the word "tabletti", (tablet in English), can appear like "tbl", "tabl", "tablettia" or "tb" and word "annos" (dose in English) can appear in different compound words like "annospussi" (dosage bag in English), "annosruisku" (syringe in English) ,"annossuihke" (dosage spray in English), "annosmitta" (measurement cup) when word "annosmitta" itself can appear in different grammatical forms like : "annosmitallinen" (a measuring cup's worth in English), "annosmitallista" (partitive form of "annosmitallinen"), "annosmittaa" (partitive from), "annospussillista" (partitive form of annospussi) and "annostelumitallinen" (portioning measurement unit in English). In many cases, we also have "dosage" and "unit name" joined together without space character in between. For example, most doctors write "1 tabletti" as "1tabletti". All these listed difficulties necessitate a robust prepossessing step before the actual labeling.

4 Model and feature extraction

This section is about model creation using CRFs and feature extraction steps.

4.1 Model

Conditional random fields (CRFs) (Lafferty et al., 2001) is a powerful method to solve labeling problem in a sequence of input word tokens. CRF models the conditional probability of a sequence of labels with respect to the input sequence. It takes into account the sequential relations between labels as well as the relations between a label and its corresponding input token. The inference is done by finding the most probable label sequence given input features, this holistic nature implies consistency, as opposed to the case where one would label each word (or n-gram) individually and separately. Here we use it to model prescription entities (dosage, dosage unit, frequency) using various linguistic and categorical features. We use Crfsuite C++ library for the implementation of our method (Okazaki, 2007). Crfsuite provides fast training and labeling and uses the standard feature templates.

4.2 Features

We make use of both linguistic as well as categorical features for the modeling problem. Table 1 lists defined feature templates that we use. Categorical features are created using two lists, first is the list of dosage unit while second is the list of frequency identifier names. Both lists are taken from a predefined list in the health care system for regular prescriptions. Naturally, these lists do not contain all possible form of dosage unit or frequencies, as we mentioned in previous section, dosages and frequencies can appear in different grammatical forms or as abbreviations or even typos.

Due to the rich morphology of the Finnish language, there is a relation between morphological categories and label output. We need the morphological analysis of prescription text to make use of this relation. In order to obtain this morphological analysis, we used Turku dependency parser (Kanerva et al., 2018). Turku dependency parser is a neural parsing pipeline for segmenta-

Identifier	Feature	Definition
F0	p_i	Current POS
F1	p_{i+1}	Next POS
F2	p_{i-1}	Previous POS
F3	c_i	Current case
F6	g_i	Current is a number, binary
F7	g_{i+1}	Next is a number, binary
F8	s_i	Current is in dosage unit list, binary
F9	s_{i+1}	Next is in dosage unit list, binary
F10	s_i	Current is in frequency list, binary
F11	r_{i+1}	Next is in frequency list
F12	r_i	Current root
F13	r_{i+1}	Next root
F14	r_{i-1}	Previous root
F15	r_{i+2}	Second next root

Table 1: List of features templates

tion, morphological tagging, dependency parsing and lemmatization for the Finnish language. We use morphological tagging outputs of Turku dependency parser in this work. There is a relation between output labels and morphemes. For example, dosage are numbers and POS tag "NUM" (number) refers to being number. Feature IsNumber is a binary feature in cases that POS tag is not "NUM" but token includes numbers like ranges. The Finnish language has very rich noun cases. Often there is a relation between the case of a token in prescription and its output label. Table 2 shows the percentage of tokens in prescriptions that have a specific case for each label. In Finnish, cases indicate the syntactic function of a noun in the sentence. The case markings are suffixed to the end of the token. Thus, the presence of a case marking in the token can give information about the label like frequency. Because frequency is mostly related to time or duration, when the token has "Adessive" case. Adessive case corresponds to prepositions "on" or "at" in English. Second informative feature for label frequency is "Inessive" which corresponds to "in" in English. Case "Allative" ("onto") has very small relation with being frequency.

5 Experimental Results

In this section, we show the experimental results for our proposed CRFs based tagging method. We tested the model using 10-fold cross-validation. In order to assess the importance of different elements of our proposed model, we train a sequence of classifiers of increasing complexity. We start with a memorization classifier, where each token is labeled individually by looking up the most frequent label it is associated with in the training data. This baseline method corresponds to a 0-order CRF with the word surface forms as the only feature. The results of this baseline classifier are shown in Table 3. Next, we try a CRF with order 1 and surface forms as features. This allows us to measure the effect of enforcing label order consistency. As seen in Table 4, the effect varies for each label, e.g. dosage labeling shows the biggest improvement over the simple memorization method. In particular, numeric tokens are hard to distinguish individually since they can be a frequency or a dosage, but when taken in the context of the token sequence they are much easier to classify. Without other more complicated features, F1 measure is over 90%, this shows that CRFs are very powerful in sequential tagging just by enforcing labeling consistency.

	Precision	recall	F1
Dosage	0.6677	0.8460	0.7464
Dosage unit	0.9562	0.9707	0.9634
Frequency	0.8361	0.9006	0.8672
Comments	0.9541	0.8337	0.8898
Macro-average	0.8535	0.8877	0.8667

Table 3: Baseline results

Case	Dosage	Dosage unit	Frequency	Comments
Adessive	0	0	84.4	15.3
Inessive	0	0	81.7	18.2
Instructive	0	0	78.5	21.4
Partitive	0	21.5	62.3	16.1
Translative	0	0	55.8	44.1
Essive	0	0.8	8.8	90.35
Genitive	0	0.7	8.5	90.63
Nominative	0	20.5	5.9	72.85
Elative	0	1.1	5.8	89.53
Illative	0	0	19	98.0
Allative	0	0	12	97.4

Table 2: Percentage of cases in labels

	Precision	recall	F1
Dosage	0.9686	0.9546	0.9616
Dosage unit	0.9642	0.9601	0.9622
Frequency	0.8909	0.9071	0.8989
Comments	0.9350	0.9303	0.9326
Macro-average	0.9396	0.9380	0.9388

Table 4: CRFs Baseline results

	Precision	recall	F1
Dosage	0.9677	0.9588	0.9632
Dosage unit	0.9733	0.9849	0.9791
Frequency	0.8951	0.9128	0.903
Comments	0.9453	0.9341	0.9397
Macro-average	0.9453	0.9476	0.9464

Table 5: Categorical features results

In Table 5 we show results of tagging for each label when we use just categorical features and the surface form of current token. As before, dosage unit benefits the most from the inclusion of categorical features.

Table 6 shows the result for tagging when we use linguistic features and surface form of the current token. F1 measure of label frequency compared to baseline and categorical feature exhibits a clear improvement. The relation between linguistic features and the tags can be observed simply by counting the associated cases. In Table 2 we show the percentage of certain grammatical cases being labeled with a given tag. It is immediately observed that most of the cases are highly informative for the labels, for example "Adessive" case strongly suggests the label frequency while eliminating the possibilities of dosage and dosage unit. On the other hand, "Translative" case is much less informative in distinguishing between a frequency and a comment; hence we require additional features and the label sequence consistency provided by CRFs to correctly identify them. It is also seen that these cases only provide negative information about the dosage label, instead, the POS tag value of "NUM" is positively associated with that label (not shown in the table).

	Precision	recall	F1
Dosage	0.9780	0.9619	0.9699
Dosage unit	0.9764	0.9806	0.9785
Frequency	0.9219	0.9444	0.9331
Comments	0.9609	0.9510	0.9559
Macro-average	0.9593	0.9594	0.9593

Table 6: Linguistic features results

In Table 7 we show results for the final model with all features. Using previous and next token information has a positive impact on F1 measure.

	Precision	recall	F1
Dosage	0.9822	0.9680	0.9751
Dosage unit	0.9819	0.9924	0.9871
Frequency	0.9253	0.9460	0.9356
Comments	0.9653	0.9542	0.9597
Macro-average	0.9636	0.9651	0.9643

Table 7: All features results

Table 8 show the accuracy for different tests. Item accuracy refers to accuracy of each token's label in prescriptions. Adding more linguistic features clearly improves the accuracy. Instance ac-

curacy is accuracy of all tokens in one prescription that are labeled correctly, i.e. even a single labeling error is counted as an error for the whole prescription. In instance accuracy we observe a remarkable improvement when we add linguistic features.

6 Post processing

In attaining this preferred state of data quality, we would be required to further classify our model results into a set of known categories found in this target information system that are defined as the subsets of natural classes of "dosage frequency" and "dosage unit", an action which we would be calling as conducting the database mapping.

For testing the accuracy of database mapping we developed an automated testing solution that would perform full end-to-end integration testing of the complete solution and simulate possible natural world usage such as concurrent and batch processing of unstructured prescriptions. The automated testing solution would use a set of 3694 hand-labeled prescriptions provided by a third-party actor as the ground-truth with guaranteed labeling accuracy of over 98% if the prescription in question had all classes labeled.

This sequential classification event creates a compound probability problem where the actual model performance can be considered as a priori probability for conducting the database mapping as its performance directly affects the results of database mapping. As a result, post-processing encounters two primary challenges: model labeling error and variance in language-specific syntax as well as semantics.

Language variance was solved by a combination of three different solutions: First we introduced internal orthography for the system by implementing robust rule-based heuristics in preprocessing that would perform spell-correction on input strings by transforming them into a more standardized language e.g. prescription string "tarv 1 1/2 -2 3/4x3 -5 pv:ssä" would be transformed into "1.5-2.75 tablettia 3-5 kertaa päivässä tarvittaessa" (In English, 1.5-2.75 tablets 3-5 times per day if required), thus reducing language complexity with negligible data loss (less than 0.5% in all categories combined). Improvement is seen in Figure 1 as iteration 2 from baseline iteration of 1.

Second, we analyzed results for string frequencies and created stemmed versions of object-relational-mapping (ORM) pair dictionaries, where the key was a stemmed class name e.g. "3 kerta päivä" and the value was in a code representation e.g. "100056" based on string occurrences. Stemming was performed on the same Turku neural parsing pipeline that is used for model generation. By matching stemmed versions of classes and model results we were able to further reduce complexity as demonstrated in Figure 1 as iteration 3.

The third solution was the implementation of approximate string matching, colloquially known as fuzzy matching, based on Levenshtein distance (Yujian and Bo, 2007) between the stemmed input string and stemmed class name. As we can see from Figure 1 iteration 4 this improved our results in "frequency" substantially. This solution had outstanding performance when the class names are relatively short e.g. unit "ml" (in English, ml, abbreviation of milliliter) compared to frequency "3 kertaa päivässä tarvittaessa" (In English, 3 times per day if required). In longer class names we experienced challenges in Hamming distance (Xu and Xia, 2011) conditions, where strings had equal length, but semantically different, class names. For example "2 times per day" and "8 times per day" have a Hamming distance of 1, but this difference has a high risk of the detrimental outcome in a clinical setting from potential under or overdose. Separability of classes was increased by writing out numbers, thus increasing their Levenshtein distance and minimizing the possible occurrences of equal length strings i.e. Hamming distance conditions.

Further on we implemented rule-based heuristics based on observed standard errors from model inference and database mapping functionality to increase our overall accuracy. This was implemented in a form of stepped funnel process, where the incorrectly mapped code representations were gathered in a list that would be processed by a set of heuristics. As a step result average of the system error would be reduced and a new list of incorrectly mapped code representations would be gathered and the process would be repeated recursively until required levels of accuracy would be attained.

For future work we will try semantic based search to solve frequency mapping problem. This can be an ontology based semantic search.

	Baseline	Categorical features	Linguistic features	All features
Item accuracy	0.9312	0.9383	0.9545	0.9588
Instance accuracy	0.6863	0.7100	0.78740	0.7998

Table 8: Item and instance accuracy for different feature sets

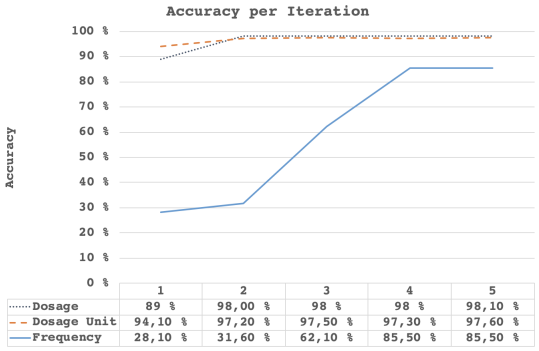

Figure 1: Mapping accuracy

7 Conclusion

In this paper, we used CRFs to model conditional probability between tokens in prescriptions and output labels, dosage, dosage unit, frequency, and comments. This model is for Finnish prescriptions. Since Finnish is an agglutinative language and has rich morphology we define two types of features. First, are categorical features which are binary features of belonging to a certain list of tokens. Second features are linguistic features which are based on the morphological analysis. In previous works, linguistic features were under-utilized. We show that linguistic features are more informative than categorical features. This model is the state of art for prescription extraction problem in the Finnish language. We are using 9692 prescriptions and our reported results are based on 10-fold cross-validation. We show that a robust pre-processing step followed by a CRF based classifier using a combination of linguistic and categorical features yield an excellent labeling accuracy. Finally by implementing heuristics in post-processing based on observed standard errors in the system we were able to reach clinical standard in classification results.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Razieh Ehsani, Muzaffer Ege Alper, Gulsen Eryigit, and Esref Adali. 2012. Disambiguating main pos tags for turkish. In *Proceedings of the 24th Conference on Computational Linguistics and Speech Processing (ROCLING 2012)*, pages 202–213.

Carol Friedman. 2000. A broad-coverage natural language processing system. In *Proceedings of the AMIA Symposium*, page 270. American Medical Informatics Association.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Zuofeng Li, Feifan Liu, Lamont Antieau, Yonggang Cao, and Hong Yu. 2010. Lancet: a high precision medication event extraction system for clinical text. *Journal of the American Medical Informatics Association*, 17(5):563–567.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Jon Patrick and Min Li. 2009. A cascade approach to extracting medication events. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 99–103.

Carson Tao, Michele Filannino, and Özlem Uzuner. 2017. Prescription extraction using crfs and word embeddings. *Journal of biomedical informatics*, 72:60–66.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Zeshui Xu and Meimei Xia. 2011. Distance and similarity measures for hesitant fuzzy sets. *Information Sciences*, 181(11):2128–2138.

Olcay Taner Yıldız, Ercan Solak, Razieh Ehsani, and Onur Görgün. 2015. Chunking in turkish with conditional random fields. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 173–184. Springer.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.