Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-146

Extracting the Population, Intervention, Comparison and Sentiment from Randomized Controlled Trials

Markus ZLABINGER [a] Linda ANDERSSON [a] Jon BRASSEY [b] Allan HANBURY [a]
[a] *Institute of Software Technology and Interactive Systems, TU Wien, Vienna*
[b] *Trip Database Ltd., UK*

Abstract. In this paper, an identification approach for the Population (e.g. patients with headache), the Intervention (e.g. aspirin) and the Comparison (e.g. vitamin C) in Randomized Controlled Trials (RCTs) is proposed. Contrary to previous approaches, the identification is done on a word level, rather than on a sentence level. Additionally, we classify the sentiment of RCTs to determine whether an Intervention is more effective than its Comparison. Two new corpora were created to evaluate both approaches. In the experiments, an average F1 score of 0.85 for the PIC identification and 0.72 for the sentiment classification was achieved.

Keywords. Information extraction, natural language processing, machine learning, sentiment analysis

1. Introduction

In the scope of the EU project KConnect[1], one task was to develop a search tool for the TRIP database[2]. This tool consists of a categorization based on medical conditions (e.g. diseases) and for each condition, appropriate treatment methods (e.g. drugs) are listed, ranked based on their effectiveness.

To generate the underlying data for this tool, we extracted information from the title and the abstract of Randomized Controlled Trials (RCTs). An RCT is a study design in which a group of people, who share a common medical condition (e.g. men with asthma), are randomly assigned to either the treatment group (e.g. treated with vitamin C) or the control group (e.g. receiving placebo). The participants of an RCT are called *Population*, the treatment used in the treatment group is the *Intervention* and the treatment of the control group is called *Comparison*. Additionally, the main outcome of an RCT, e.g. is vitamin C more effective than placebo, is reported.

In this paper, we propose an automatic identification approach for the Population, the Intervention and the Comparison (short PIC) in RCTs (Section 2). In previous approaches [1,2,3] the PIC identification was done on a sentence level; however, we propose a precise word level identification based on hand-crafted rules. Furthermore, we propose a sentiment classification method based on machine learning, to determine if the Intervention of an RCT is more effective than its Comparison (Section 3). To evaluate both methods, we created two new corpora with the help of human annotators. The evaluation results for these corpora are given in Section 4.

[1] EU project about search in the medical domain: http://www.kconnect.eu/
[2] A clinical search engine: http://www.tripdatabase.com

2. Method for the PIC Identification

In previous approaches, sentences were classified to determine if they contain a certain PIC element, or not. Only [4] proposed a word level approach. Unfortunately, since no appropriate dataset existed (at that time), no evaluation results were provided.

The evaluation data used in our approach was generated by six human annotators (2 linguists and 4 persons from the medical domain). We developed a web-interface (Figure 1), which was used to submit annotations. To evaluate the performance of individual annotators, we created a ground truth based on 20 RCTs in cooperation with a medical expert. These 20 RCTs were annotated by each annotator. As evaluation metric, we define the agreement as $\frac{\#\text{Correct}}{\#\text{Correct}+\#\text{Incorrect}}$, where an annotation was counted as $Correct$ if it was exactly the same as the ground truth annotation and $Incorrect$ in all other cases. Stop words (e.g. *in*, *to*, *[*, *we*) were removed before computation of the agreement. For titles, on average, the agreement was 0.70 for P, 0.66 for I and 0.62 for C. We also asked the users to submit annotations for abstracts; however, due to a more complex sentence structure (e.g. more text variety, longer sentences), we reached agreements of less than 0.50 for each PIC element.

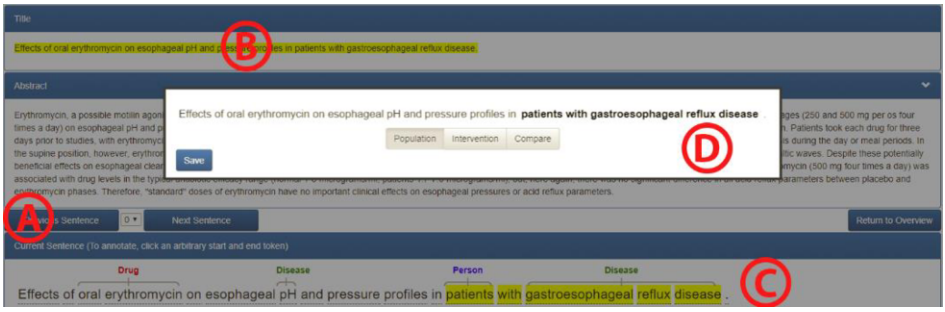

Figure 1. The PIC annotation web-interface: (A) Sentence navigation, (B) active sentence (yellow background), (C) active sentence split into single word units (tokens) and finally, after selecting a start and end token, a pop-up window (D) is shown and used to submit an annotation for either P, I or C.

Based on the newly created corpus, we propose an automatic approach for the PIC identification. Since the annotation agreement for abstracts was weak, we decided to focus on titles of RCTs. In fact, the title already contains most of the PIC information: The 20 ground truth RCTs showed a coverage of 18/20 for P, 19/20 for I and 7/20 for C. Usually, no-medication and placebo comparisons are omitted in the titles, which explains the low 35% coverage for C. Therefore, if no C was detected, we assumed *no-medication*.

To identify PIC elements automatically, we propose a rule-based approach. Rules are hand-crafted expressions and are used to exploit commonly occurring linguistic patterns. There are frameworks that ease the rule-crafting process. In our approach, we used Stanford's TokensRegex [5], which is best explained based on an example: Assume, we want to identify the P element in *The bioavailability of nasogastric versus trovafloxacin in healthy subjects*. First, we use CoreNLP [6], a natural language processing toolkit, to split the sentence into tokens (=single word units) and for each token, the lemma (base form) and the part-of-speech (POS) tag is computed. Additionally, we use GATE's BioYodie[3], a pipeline to identify medical semantics (e.g. drugs, diseases),

[3] https://gate.ac.uk/applications/bio-yodie.html

and a static lookup list, which contains 42 person keywords (e.g. children, asthmatics), to add semantic information to the tokens. The resulting token representation is shown in Figure 2. Finally, we apply the TokensRegex rule: **[word:of] []* [word:in] ($Population [pos:JJ]* [sem:/Person/ or sem:/Disease/])**, which consists of following components:

- **[word:of]:** The first token must be *of* on the word layer. (Token3 in Fig. 2)
- **[]* [word:in]:** The *of* token is followed by any token (=[]) zero or more times (=*); however, eventually the word layer token *in* must occur. (Tokens 4 to 7)
- **[pos:JJ]*:** After the *in* token, zero or more adjective tokens may occur (JJ = adjective), i.e. the rule would also fire for ... *[tall energetic] subjects*. (Token8)
- **[sem:/Person/ or sem:/Disease/]:** The end token must be a *Person* or *Disease* on the semantic layer; i.e., ... *in severe [headache]* would also match. (Token9)
- **($Population ...):** Round brackets mark a capture group, i.e. non-Population tokens (e.g. *of*) are not captured and excluded in the result set. (Tokens 8 to 9)

5:				Drug		Drug			Person
4:	DT	NN	IN	NN	CC	NN	IN	JJ	NNS
3:	the	bioavailability	of	nasogastric	versus	trovafloxacin	in	healthy	subject
2:	The	bioavailability	of	nasogastric	versus	trovafloxacin	in	healthy	subjects
1:	**Token1**	**Token2**	**Token3**	**Token4**	**Token5**	**Token6**	**Token7**	**Token8**	**Token9**

Figure 2. Token representation: (Layer 1) index, (2) raw word, (3) lemma, (4) POS tag[4] and (5) semantics.

When crafting the rules, we did not differentiate between C and I; because only after identifying the full Intervention/Comparison phrase, we separated both elements based on keywords (e.g. *vs.* or *compared to*). We also considered trigger words, i.e. words that are not part of the Intervention/Comparison; but, may influence their separation. For example, consider the annotated sentence *Efficacy of **I[aspirin with vitamin c]** and **C[placebo]** in **P[men]***. If we replace the text *Efficacy of* with a trigger word, e.g. *Comparing **I[aspirin]** with **C[vitamin c and placebo]** in **P[men]***, the I and C would change.

3. Method for the Sentiment Analysis

We created the sentiment corpus in a similar way as the PIC corpus: Six annotators were asked to submit annotations through a web-interface (Figure 3) and individual agreements were computed based on a 30 RCT ground truth (again, in coop. with a medical expert). The annotators were instructed to select a sentiment (positive, neutral or negative) for conclusion sentences in the abstracts of RCTs. In a structured abstract (=with headings), sentences within the *Conclusion* heading were considered as conclusion sentences. For unstructured abstracts (=no headings), we computed the number of conclusion sentences as $MAX(0.125 \times \text{totalSentences}, 1)$; e.g., for an abstract with 13 sentences, the last $MAX(1.625, 1) \approx 2$; sentences are considered as conclusion. The relative value of 0.125 was computed based on the analysis of 2000 structured abstracts.

When evaluating the annotations of individual users, we observed a rare occurrence ($\sim 10\%$) of *negative* samples. Based on the small number of *negative* samples, we could

[4] DT=Determiner, NN=Noun, IN=Preposition, CC=Conjunction, JJ=Adjective and NNS=Noun Plural

not create a machine learning model that also works well for the *negative* class. Therefore, we merged the *negative* and the *neutral* class. After merging, we computed the annotation agreement (defined in Section 2) for each user, which was on average 0.78.

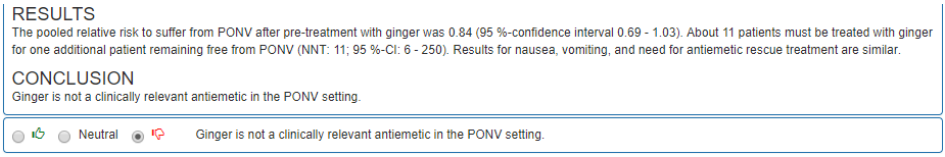

Figure 3. The sentiment annotation web-interface.

Based on the newly created sentiment corpus, we created a machine learning model. For this, we evaluated various configurations that also performed well in other classification approaches in the medical domain (e.g. [7,3]). In detail, as text input, we used raw text, lemmas (children with headache were → child with headache be), or POS tags (children → children_NNS). As textual features, uni-gram (Child, with, ...), bi-gram (Child with, with headache, ...), three-gram and all possible combinations of the listed features were evaluated. As classifiers, we evaluated Logistic Regression, Support Vector Machine (SVM), Random Forest, Multinomial Naive Bayes, and Bernoulli Naive Bayes.

4. Results

PIC identification: To create the PIC corpus, we asked each annotator to annotate 500 RCTs; for which, 250 were identical (i.e. overlapping) for all annotators. Based on a majority voting for the overlapping annotations, we created an evaluation dataset. This dataset showed agreements of P 0.89, I 0.84 and C 0.71 when compared to the expert annotations (i.e. the 20 RCTs) and contained 217 P, 220 I and 76 C elements.

We created 7 rules for the Population and 14 rules for the Intervention/Comparison. We measured the effectiveness of our approach on a test set (20% of the dataset) on a token-level. The idea of a token-level evaluation, i.e. true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), is described based on an example in Figure 4. The evaluation results for the test set are shown in Table 1.

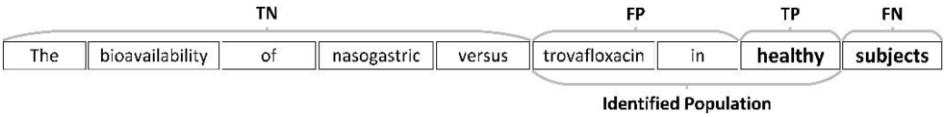

Figure 4. Token-level evaluation for the Population *healthy subjects* versus the identified Population *trovafloxacin in healthy*: (TN) correctly ignored non-P tokens, (FN) missed P tokens, (TP) correctly identified P tokens and (FP) incorrectly identified non-P tokens

Identification errors usually occurred in sentences that did not conform to common linguistic patterns or if the separation process for the Intervention/Comparison was complicated (e.g. ***I[C****[oxycodone]] alone and* ***I****[combined with ethanol]*). Additionally, Populations consisting of several prepositional connectors, e.g. *men with [...] from [...] after experiencing [...]*, were problematic.

Table 1. Rule-based PIC identification results for the test set.

PIC Type	TP	FP	FN	Precision	Recall	F1
Population	121	15	18	0.89	0.87	0.88
Intervention	84	33	11	0.72	0.88	0.79
Comparison	20	3	3	0.87	0.87	0.87

Sentiment Analysis: As dataset for the sentiment analysis, we used the annotations of the two best performing annotators, who reached an agreement of about 0.8 for the sentiment, when compared to the expert annotations (i.e. the 30 RCTs).

The sentiment dataset consisted of 619 neutral and 532 positive sentences. The best result, with an F1 score of 0.72 (precision 0.76, recall 0.69), was achieved when using lemma preprocessing, uni-gram features and an SVM classifier (hyper parameters: C=1, gamma=1 and kernel=rbf).

Classification errors occurred for sentences that state an increase or decrease of certain medical parameters (e.g. *ammonia levels tended to decrease*). Now, it is unclear, at least for the classifier, if an increase of ammonia is positive or not

5. Conclusion

In this paper, we showed that with a few hand-crafted rules it is possible to identify the Population, Intervention and Comparison in the titles of RCTs with an average F1 score of 0.85. Furthermore, we showed that the sentiment of an RCT can be predicted based on the sentences of the abstract with an F1 score of 0.72.

In future research, we plan to create more advanced models for the sentiment analysis and second, incorporate abstracts for the PIC identification.

References

[1] F. Boudin, J.-Y. Nie, J. C. Bartlett, R. Grad, P. Pluye, and M. Dawes, "Combining classifiers for robust PICO element detection," BMC Medical Informatics and Decision Making, vol. 10, no. 1, p. 29, 2010.A.N. Author, Article title, Journal Title 66 (1993), 856–890.

[2] S. N. Kim, D. Martinez, L. Cavedon, and L. Yencken, "Automatic classification of sentences to support evidence based medicine," BMC Bioinformatics, vol. 12, no. 2, p. S5, 2011.

[3] H. Hassanzadeh, T. Groza, and J. Hunter, "Identifying scientific artefacts in biomedical literature: The evidence based medicine use case," Journal of Biomedical Informatics, vol. 49, pp. 159 – 170, 2014.

[4] S. Chabou and M. Iglewski, "PICO extraction by combining the robustness of machine-learning methods with the rule-based methods," in 2015 World Congress on Information Technology and Computer Applications (WCITCA), 2015, pp. 1–4.

[5] A. X. Chang and C. D. Manning, "TokensRegex: Defining cascaded regular expressions over tokens," Department of Computer Science, Stanford University, Tech. Rep. CSTR 2014-02, 2014.

[6] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in Association for Computational Linguistics (ACL) System Demonstrations, 2014, pp. 55–60.

[7] Y. Niu, X. Zhu, J. Li, and G. Hirst, "Analysis of polarity information in medical text," in AMIA Annual Symposium Proceedings, vol. 2005. American Medical Informatics Association, 2005, pp. 570–574.