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

*doi:10.3233/978-1-61499-852-5-680*

# Unsupervised Information Extraction from Behaviour Change Literature

Debasis GANGULY[a], Léa A. DELERIS[a], Pol MAC AONGHUSA[a],
Alison J. WRIGHT[b], Ailbhe N. FINNERTY[b], Emma NORRIS[b],
Marta M. MARQUES[b], Susan MICHIE[b]
[a] *IBM Research, Dublin, Ireland*
[b] *Centre for Behaviour Change, University College London, UK*

**Abstract.** This paper describes our approach to construct a scalable system for unsupervised information extraction from the behaviour change intervention literature. Due to the many different types of attribute to be extracted, we adopt a passage retrieval based framework that provides the most likely value for an attribute. Our proposed method is capable of addressing variable length passage sizes and different validation criteria for the extracted values corresponding to each attribute to be found. We evaluate our approach by constructing a manually annotated ground-truth from a set of 50 research papers with reported studies on smoking cessation.

**Keywords.** Behavior Change, Smoking Cessation, Information Extraction

## 1. Introduction

Behavioural change interventions (BCIs) are policies, activities, services or products designed to cause people to act differently from how they would have done otherwise. They involve attempting to change either members of the target population (in terms of their knowledge, skills, feelings or habits), or their social or physical environment.

Research findings have the potential to provide invaluable knowledge to help with developing or selecting BCIs but this evidence needs to be synthesised and interpreted [4]. Since the scientific literature on behaviour change is vast and accumulating at a rapidly accelerating rate, it is difficult to achieve this manually. This necessitates the development of automatic information extraction (IE) approaches to construct a knowledge base of behaviour change findings. An automated IE approach that extracts relevant pieces of information from BCI reports can act as a first step to design navigable interfaces that allow domain experts to easily find relevant pieces of information from previously reported studies. The extracted information can also be used as features to develop predictive models of outcomes for BCI thought experiments.

IE approaches are typically supervised in nature. Sequence models, such as conditional random field, have been applied to extract information from unstructured text [2, 1]. Such supervised approaches mainly rely on the availability of manually annotated data to train the IE models. In contrast, this paper describes our work towards unsupervised information extraction (IE) from a collection of BCI evaluation studies.

The main advantage of an unsupervised approach is that it does not rely on the availability of a labelled training data. The labelled data needed for an unsupervised method is for evaluation only. We make use of a manually annotated ground-truth from a set of 50 research papers with reported studies on smoking cessation. Our unsupervised

IE approach is based on a general passage retrieval framework for extracting a wide range of attribute values, ranging from characteristics of the study subjects to behaviour change techniques (BCTs).

## 2. Passage Retrieval based IE Approach

We approach the problem from an information retrieval (IR) point of view. We formulate a query for each information unit that we want to extract to obtain a list of passages ranked in descending order by their similarities with the query. To extract the answer value, we first make use of a validation criterion to filter out the likely answer candidates. We then score the candidate answers by a term proximity model that takes into account the differences in position between the query terms and the candidate answers.

We now describe each component of our IE framework in more details.

**Indexing and Information Units.** Text from each document is extracted and stored into separate fields such as 'Introduction', 'Content', 'Table body', 'References' etc. Such a field based representation of documents in the index allows application of field-based retrieval models [5, 6]. Field-based representation provides the flexibility of incorporating prior beliefs in the passage scoring function. For example, the passage scoring function can consider the fact that a valid answer candidate found in the 'references' section is less likely to be correct than those found in the 'experiments' section of a paper.

For each attribute value to be extracted, we define an information unit (IU) comprised of: i) a type, ii) a query, iii) a validation criterion for the answer (e.g. the value of the *average age* attribute must be numerical) and iv) a threshold cosine similarity value between the query and retrieved passages. An IU can be either of type: i) *value extraction* (VE), where the system aims to extract a value of an attribute, e.g. the average age of the participants in the BCI study; or ii) *detect presence* (DP), where the system predicts whether there exists enough evidence in a reported study to suggest the presence of an attribute, e.g. the whether a study prescribed *self monitoring of behaviour* for the participants. The VE type IUs are associated with a validation criterion function, whereas the DP type ones are associated with a threshold similarity value.

Equation 1 represents answer validation criterion function associated with a VE-type IU which assigns a value of 1 to a candidate answer term $a \in V$ ($V$ being the vocabulary) if $a$ satisfies some constraint, e.g. $a$ is numerical.

$$\phi(a): V \rightarrow \{0,1\} \tag{1}$$

Equation 2 denotes the similarity threshold function associated with a DP-type IU, indicating that the function evaluates to 1 if the similarity between a retrieved passage and a query, denoted by $sim(P, Q)$, is higher than a threshold $\tau$.

$$\begin{aligned} \psi(P, Q, \tau) &= 1, if\ sim(P, Q) > \tau, \tau \epsilon [0,1] \\ &= 0, otherwise \end{aligned} \tag{2}$$

**Proximity-based Ranking Function.** Each query in an IU can range from simple factoid seeking type, e.g. 'the minimum age of the participants' to more complex types, where the information resides in an arbitrarily long passage, e.g. 'the follow up treatment

after intervention'. The queries are structured in nature with Boolean operators connecting the constituent terms. As a particular example, the query for extracting the 'age' attribute is '*participant AND (age OR year OR old)*', i.e. we are interested to retrieve passages that must contain the word 'participant' and should also have one or more of the words - 'age', 'year' or 'old'.

Given a query, the retrievable units comprise arbitrary passages of text. The system constructs an in-memory transient index of passages of text while processing each document in turn. A passage in this case is comprised of a fixed length window of $w$ words. To account for different granularity in the range of contextual evidences, we use different values of $w$ to define different retrievable units. In our experiments, we set $w$ to 5, 10, 20 and 30 words. The intention of retrieving passages is to restrict extraction of factoid answers to potentially relevant small semantic units of text rather than the text of the whole document. This passage based retrieval also ensures that the proximity of the answer terms to the query terms is taken into account.

The position of the candidate answer term with respect to the query terms can potentially be useful to predict the relevance of a retrieved passage. Consequently, the ranking function needs to consider the relative distances between the positions of the query terms and the candidate answer terms. Equation 3 formally describes the proximity based ranking function between a passage $P$ and a query $Q$, denoted by $sim(P, Q)$.

$$sim(P,Q) = \frac{1}{|Q|}\sum_{q\in Q}\sum_{a\in A}\exp\left(-\frac{(p_a-p_q)^2}{\sigma}\right),\ A = \{a: \phi(a) = 1\} \quad (3)$$

The set A denotes the set of candidate answer terms, i.e. those terms for which the validation criterion function $\phi$ returns 1. Practically, for each word in the passage that matches the query terms ($q$), the similarity function increases the score for passage by an amount that depends on the distance between that word and the candidate answer ($p_a - p_q$) . Specifically, we use a Gaussian function centered at each query term to determine the increase in similarity score. The parameter $\sigma$ controls the bandwidth of the Gaussians and is set to 1 in our experiments. Such term proximity based language models for ranking documents have been proposed in [7, 3]. The main difference between these approaches and our work is that our similarity function aggregates the positional differences for only the candidate answer terms instead of aggregating this over each term as in [7, 3].

To illustrate how the similarity function of Equation 3 works in practice, we consider the following four passages retrieved from a document in our dataset with the query defined for the 'age' attribute, i.e. the query '*participant AND (age OR year OR old)*'.

Passage-1: ...avoided by smokers quitting before **age** *30* **years**...

Passage-2: We enrolled **participants aged** *18* **years** and **older**...

Passage-3: **Age** of smoking initiation (**years**)...

Passage-4: ...*3* additional **years** of life for every *100 40*-**year-old** smokers...

It can be seen that there are multiple places in the text where query terms occur (shown in bold). Passage-2 contains the relevant piece of information that needs to be extracted, suggesting that the age of the participants was 18 and over. The key observation is to note the number of query terms found in a passage and the differences in positions between the candidate answer term (which in this case is an integer number shown in italics) and those of the query terms. The lower this number is, the better is the likelihood of the passage to be relevant as modeled by Equation 3. For example, in passage-2, we can find 4 query terms at positions -2, -1, 1 and 3 relative to the candidate

answer term (the number 18). Passage-4, on the other hand, has 3 candidate answer terms and 3 query terms. The query terms 'years', 'year' and 'old' are placed -6, 1 and 2 positions apart relative to the candidate answer term 40. It is easy to see that the sum of Gaussians, centered at the positions of the candidate answer terms (Equation 3), assigns a higher score to passage-2 as compared to passage-4.

**Table 1.** Information extraction effectiveness of different IUs in our experiments

| Information to Seek | Type | Query Representation | Criteria/ Threshold | Accuracy |
|---|---|---|---|---|
| Minimum Age | VE | Participant AND (age OR year OR old) | Integer | 0.31 |
| Maximum Age | VE | Participant AND (age OR year OR old) | Integer | 0.12 |
| Average Age | VE | Average OR mean) AND (age OR year OR old) | Numerical | 0.46 |
| Gender | VE | Male OR female OR gender | 'male','female' | 0.32 |
| Average Accuracy for Value Extraction Information Units | | | | 0.30 |
| Goal Setting(Behaviour) | DP | (goal OR target) AND (quit OR plan) | 0.25 | 0.74 |
| Problem Solving | DP | cope overcome identify problem relapse | 0.25 | 0.62 |
| Action Planning | DP | action plan intention quit | 0.25 | 0.64 |
| Feedback on behaviour | DP | patient feedback | 0.25 | 0.50 |
| Self-monitoring of behaviour | DP | self monitor diary track | 0.25 | 0.88 |
| Social support (unspecified) | DP | quit instruction advice training | 0.25 | 0.50 |
| Information about health consequences | DP | hazard smoking | 0.25 | 0.60 |
| Information about social and environmental consequences | DP | harmful chemical environmental consequences | 0.25 | 0.82 |
| Pharmacological support | DP | nicotine gum patch NRT transdermal | 0.25 | 0.86 |
| Reduce negative emotions | DP | negative emotion stress | 0.25 | 0.82 |
| Average Accuracy for Detect Presence Information nits | | | | 0.698 |

## 3. Evaluation

**Dataset.** The dataset used for evaluation is composed of a set of 50 published papers on BCI studies. The papers were selected by a team of 4 domain experts. Annotation corresponding to each IU for a particular document was performed by two human annotators using the EPPI tool[1]. Conflicts were resolved through discussions. The annotation process involved highlighting relevant pieces of text. For VE-type attributes, the highlighted text comprises the answer value, e.g. the value of average age. For DP-type attributes, the highlighted text comprises evidence in the text which supports a given claim, e.g., the highlighted text 'set a target quit date' provides evidence to the DP attribute 'Goal setting (behaviour)'.

[1] http://eppi.ioe.ac.uk/CMS/Default.aspx?alias=eppi.ioe.ac.uk/cms/er4&

**Results.** Table 1 shows the average accuracy values measured per IU across the collection of 50 documents. For VE-type IUs, we consider the extracted answer to be correct if it matches exactly the ground-truth answer among papers with an annotation. In other words, the accuracy for VE type is measured as the ratio of the number of papers with correct prediction divided by the number of papers annotated with the attribute. For DP-type IUs, the system prediction is a Boolean value based on the similarity threshold (see Equation 2). The predicted answer is considered to be correct if the ground-truth contains an annotation for this attribute. We see that the overall accuracy is satisfactory for DP-types. While average accuracy for VE-type is around 30%, this should be understood as the baseline performance of our system.

## 4. Conclusions and Future Work

This paper presents initial research direction towards development of an unsupervised IE system for extraction of relevant features from BCI reports. We proposed a passage retrieval based approach that uses a combination of a term-proximity based model, answer validation criterion and a similarity threshold for extracting attribute values from relevant passages in BCI reports. Experiments conducted on a set of 50 documents show that the proposed approach yields adequate baseline effectiveness, with average accuracy of about 58%. In future, we would like to explore ways of dynamically setting the similarity threshold values for different DP-type attributes to further improve results. We also intend to explore ways of automatically formulating queries for the IUs based on the context of manually highlighted text from the documents.

## Acknowledgements

This work was supported by a Wellcome Trust collaborative award [The Human Behaviour-Change Project: Building the science of behaviour change for complex intervention development, 201,524/Z/16/Z].

## References

[1] L. A. Deleris and C. Jochim. Probability statements extraction with constrained conditional random fields. *Studies in health technology and informatics*, **228** (2016), 527-531.

[2] R. Gupta and S. Sarawagi. Creating probabilistic databases from information extraction models. *VLDB'06* (2006), 965–976.

[3] J. Leveling, D. Ganguly, S. Dandapat, and G. J. F. Jones. Approximate sentence retrieval for scalable and efficient example-based machine translation. *COLING'12 (2012),* 1571–1586.

[4] S. Michie, J. Thomas, M. Johnston, P. Mac Aonghusa, J. Shawe-Taylor, M. P. Kelly, L. A. Deleris, A. N. Finnerty, M. M. Marques, E. Norris, et al. The human behaviour change project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation. *Implementation Science* **12** (2017) 121.

[5] P. Ogilvie and J. Callan. Combining document representations for known-item search. *SIGIR '03* (2003),143–150.

[6] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. *CIKM* '04(2004),42–49.

[7] J. Zhao and Y. Yun. A proximity language model for information retrieval. *SIGIR'09* (2009), 291–298.