Risk Event and Probability Extraction for Modeling Medical Risks

Charles Jochim and **Bogdan Sacaleanu** and **Léa A. Deleris**
IBM Research – Dublin, Ireland
[charlesj—bogsacal]@ie.ibm.com

Abstract

In this paper we address the task of extracting risk events and probabilities from free text, focusing in particular on the biomedical domain. While our initial motivation is to enable the determination of the parameters of a Bayesian belief network, our approach is not specific to that use case. We are the first to investigate this task as a sequence tagging problem where we label spans of text as events A or B that are then used to construct probability statements of the form $P(A|B) = x$. We show that our approach significantly outperforms an entity extraction baseline on a new annotated medical risk event corpus. We also explore semi-supervised methods that lead to modest improvement, encouraging further work in this direction.

1 Introduction

Probabilistic risk analysis has long been applied to the biomedical domain for either punctual decision support (Deleris et al. 2006; Pietzsch and Pat-Cornell 2008), risk prediction and detection (Watt and Bui 2008; Steele et al. 2012), or as the inference engine within a decision support system (Warner et al. 1988; Fuller 2005; Hoffer et al. 2005). Bayesian belief networks (BBNs) (Pearl 1988) are a popular underlying modeling framework for such analysis. A simple BBN of only three nodes can be seen in Figure 1. A BBN is composed of (i) a structural layer (the nodes and arcs linking them) that depicts the variables and their associated dependence and independence relationships and of (ii) a quantitative layer, typically captured in the form of conditional probability tables (CPTs), which are composed of statements of the form $P(C = \textit{true}|A = \textit{true}, B = \textit{false}) = 0.01$. The focus of the present paper is on extracting information to better evaluate the quantitative layer, i.e., the parameters of the CPTs. When sample data is not available, the traditional method to evaluate those parameters has been to rely on expert elicitation. In fact, a large body of research has focused on designing methods to ensure high quality elicitation (Cooke and Goossens 1999). However, such approaches are time consuming (e.g., 10 hours were spent to evaluate the 900 probabilities of a 39-variable model as reported in (van der Gaag et al. 2002)); costly (as they require experts' and analysts' time); and finally, cognitively difficult for the experts. In addition, the manual effort spent eliciting probabilities can usually not be used in another context or even updated as newer medical knowledge is acquired. Ideally, we would like to automatically build the quantitative layer of a BBN, assuming the structural layer has been provided separately, but ensure that we are still using high-quality expert information. Our solution for this is to (i) extract relevant medical quantitative risk information from published (medical) literature, (ii) normalize the extracted events to match the variables in the BBN structure provided, and finally (iii) use the quantitative information extracted to determine the quantitative parameters of the network. In this paper we concentrate on the first step, leveraging NLP techniques for the extraction of these medical risk events.

The rest of the paper is structured as follows. In Section 2, we further describe the risk event extraction task and its associated challenges. Our approach for tackling this problem is covered in Section 3. Then in Section 4 we present our results along with some discussion. This paper relates to work in several different areas of the literature, which are reviewed in Section 5. Finally, we conclude in Section 6.

2 Task Description

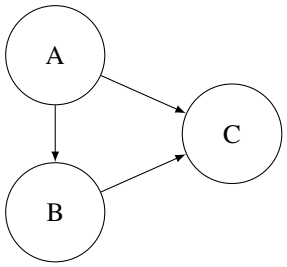
2.1 Background

Our focus is on the extraction of quantitative risk information in the form of probability statements. In this section, we describe the probability terms (numbers) that we extract along with the events A (the conditioned event) and B (the conditioning event) that are part of the conditional statement. *Event extraction* is an active area of BioNLP research (Nédellec et al. 2013) that identifies bio-molecular events on proteins and genes, and we must distinguish our *risk* event extraction from that body of work. Bio-events from the BioNLP shared task (e.g., gene expression or binding) define the relations between different entities (e.g., genes, proteins, or RNA types). Instead, we use "event" in the probabilistic sense, which means that risk event detection is more similar to entity detection in BioNLP research.

Risk events are quite heterogeneous and exhibit a broader semantic variety than the bio-molecular entities. "1977-1990", "breast cancer", "homozygous carriers", "> or = 10 years", and "younger than age 35" are all examples of different events taken from our corpus. This variety owes mainly

A	
true	*false*
0.2	0.8

B	A	
	true	*false*
false	0.4	0.6
true	0.01	0.99



A	B	C	
		true	*false*
false	*false*	0.4	0.6
false	*true*	0.01	0.99
true	*false*	0.01	0.99
true	*true*	0.01	0.99

Figure 1: Example of Bayesian belief network with three variables

Odds	$\text{odds} = P(A)/P(\neg A)$
Odds ratio	$\text{OR} = \frac{\frac{P(A\|B)}{P(\neg A\|B)}}{\frac{P(A\|\neg B)}{P(\neg A\|\neg B)}}$
Relative risk	$\text{RR} = \frac{P(A\|B)}{P(A\|\neg B)}$
Hazard ratio	$\text{HR} = \frac{P(A(t)\|B)}{P(A(t)\|\neg B)}$
Population attributable risk	$\text{PAR} = P(A\|B) - P(A)$

Table 1: Probability terms used in medical journals

to the fact that only the explicitly mentioned part of the event has been annotated. We will need to develop new techniques for detecting ellipses, metonymy, and anaphora resolution to obtain more complete event definitions. In this initial work, we focus on common risk events with a stronger signal in the corpus, such as "disease-free survival" or "breast cancer incidence". Those will eventually constitute the variables that appear in the BBN.

Probability expressions can take a number of forms. Even the simplest probability statements can be complicated in free text, represented as different combinations of words and numbers, e.g., "Thirty-six percent", "78.2 percent", "51.7%", "5/1124", or "51 out of 118". Two of the most common forms of probability statements are odds ratio (OR) and relative risk (RR). In probability, the odds of an event is defined as the probability that the event will occur divided by the probability that the event will not occur. Odds ratios are ratio of odds for different subgroups. For instance, if we were to compare nulliparous women (no live birth) with women having at least one live birth (mono/multiparous) for the risk of developing breast cancer, then the odds ratio would be:

$$\text{OR} = \frac{\frac{P(\text{breast cancer}|\text{nulliparous})}{P(\text{no breast cancer}|\text{nulliparous})}}{\frac{P(\text{breast cancer}|\text{mono-/multiparous})}{P(\text{no breast cancer}|\text{mono-/multiparous})}}$$

Table 1 summarizes some of the more common probability terms found in the medical literature and their equivalent probability equations. As in our example, variables A, B, and $\neg B$ (the complement of B) could be replaced with values "breast cancer", "nulliparous", and "mono-/multiparous" respectively.

Along with the probability terms we would like to identify and extract the events A and B from text. Risk event identification presents multiple challenges. Consider the example "Carriers of the AC haplotype, which represents the variant alleles of both SNPs, were at an increased risk (OR = 1.41, 95% CI 1.09-1.82)." We have an odds ratio probability term "OR = 1.41" and two events. However, determining the boundaries of the events is not straightforward. Should the conditioning event be the whole subject including the relative clause, only "Carriers of the AC haplotype," or even simply "AC haplotype"? Another problem illustrated by this example is that from this limited context we do not know that "risk" refers to risk of breast cancer. Anaphora resolution should be added to the pipeline by linking "risk" to "breast cancer" before construction of the BBN. This is left for future work.

There are other problems not found in the above example. In some cases one of the events may not be explicitly expressed in the sentence and must be inferred by the reader. Finally, although it is usually clear to a human reader, it can be challenging to determine automatically which is the conditioned event and which is the conditioning event. The distinction, however, is essential from a probability perspective, as typically $P(A|B) \neq P(B|A)$. Consider for instance the case where A represents *Winning the lottery* and B represents *Playing the lottery*.

After identifying probability terms and extracting the related events, we can construct the conditional statements, which after some post-processing will be used to evaluate the parameters of a BBN. To better understand the task, we annotated a corpus of probabilities and risk events. We cover some details of how this corpus was created in the next section.

2.2 Corpus

Our risk event probability corpus is comprised of 200 abstracts returned from PubMed that match the query "breast cancer AND parity". This follows in the tradition of other biomedical corpora: GENIA (Kim et al. 2003), for example, started as a set of abstracts from PubMed, which contained the terms "human," "blood cell," and "transcription factor."

Some statistics of the corpus can be seen in Table 2. Probability and event annotation was conducted in two rounds. For the first round, three annotators independently annotated each sentence in the corpus as a probability sentence or not

Abstracts	200
Sentences	2045
w/ events or probability	376
w/ event A	315
w/ event B	316
Avg. sentence/abstract	10.2
Avg. sentence length (tokens)	37.9
Avg. event/abstract	4.5
Avg. length event A (tokens)	2.3
Avg. length event B (tokens)	2.3
Avg. event A/prob. sentence	1.1
Avg. event B/prob. sentence	1.3

Table 2: Risk Event Corpus statistics

(i.e., containing at least one probability expression) and then for each probability sentence the events A and B were annotated. We measure inter-annotator agreement for probability sentence detection by averaging Cohen's kappa, $\kappa = 0.920$. For the second round of annotation, a risk expert with experience assessing risk in the biomedical domain resolved annotation disagreements and corrected the event boundaries when they differed between annotators. Although, the annotators are not experts in the biomedical domain, this type of annotation does not require medical expertise. The goal of the annotation is to identify events in a conditional statement and this task is fairly domain independent. For instance, annotators can identify events A and B in "Average duration of breast-feeding of 11-12 months reduced risk of breast cancer by 54% compared with the duration of 1-4 months" and "At least 75 percent of all hydraulic systems fail due to contaminated or aging hydraulic fluid" without being experts in either medicine or mechanical engineering.

3 Labeling Risk Events

3.1 Choice of Algorithm

Our goal is to accurately identify the probability terms and events in conditional statements in medical texts. As we normalize event text before using them for probability aggregation, it is not highly critical to extract the exact boundary. For example, it would still be acceptable in our system that only "breast cancer" would be detected when the annotated event is "development of breast cancer" (or vice versa).

We chose to explore whether sequence labeling could be relevant for risk event identification. In particular we use conditional random fields (CRF) (Lafferty, McCallum, and Pereira 2001) which have been shown to work well for related NLP tasks (Sha and Pereira 2003; Settles 2004). We use a linear-chain CRF defined as:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}} \exp(\sum_{t=1}^{T} \sum_{k} \theta_k f_k(y_{t-1}, y_t, \mathbf{x}_t))$$

where $Z_{\mathbf{x}}$ normalizes over the entire input sequence (e.g., sentence), $Z_{\mathbf{x}} = \sum_y \exp(\sum_{t=1}^{T} \sum_k \theta_k f_k(y_{t-1}, y_t, \mathbf{x}_t))$. With a first-order CRF, features can be defined on the pair of output tags y_{t-1} and y_t.

word	dependency label*
lemma	head POS*
POS tag	predicate*
position	distance to predicate*
word shape	

Table 3: Feature list. Features marked with an asterisk (*) were extracted but did not improve classification and were removed in feature selection.

3.2 Features

Our experiments use standard sequence tagging features which are summarized in Table 3. We look at features at different levels of linguistic complexity – from simple surface forms to dependence relations from a parse tree.

Features are extracted for each token, t, that is to be tagged with the CRF. Our first features are simply the surface form of the word (w_t), lemma ($lemma_t$), and the part-of-speech tag (pos_t). The word shape feature ($shape_t$) is extracted using Stanford CoreNLP[1] with the default word shape algorithm similar to the one in Bikel et al. (1997).

We explored deeper linguistic features taken from dependency-based parsing. For example, we use the arc label pointing to t and the POS tag of t's head. Using the dependency parse we also retrieve the predicate of the sentence and include a feature measuring the distance to the predicate via dependency arcs. These features did not help classification on the development set and are not included in the experiments in Section 4. In preliminary experiments we used a semantic feature – a UMLS tag – obtained via MetaMap[2] (Aronson 2001). This is a potentially useful feature but is expensive to extract (cf. Wang and Fan 2014), and does not scale well with the semi-supervised solutions we propose later in Section 4.1.

One of the advantages of using a sequence tagger like CRF is that it utilizes the features from adjacent tokens. The linear-chain CRF we use defines feature templates for zeroth-order features and first-order features. Not all of the features listed above are effective as context features, and not all are useful as first-order features. We start by looking at the following context features: words ($w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$), lemmas ($lemma_{t-2}, lemma_{t-1}, lemma_{t+1}, lemma_{t+2}$), part of speech ($pos_{t-2}, pos_{t-1}, pos_{t+1}, pos_{t+2}$), and word shape ($shape_{t-2}, shape_{t-1}, shape_{t+1}, shape_{t+2}$), and then use feature selection to choose the optimal amount of context to the left and right for each feature.

4 Results

4.1 Experimental Setup

We take all of the sentences which contain any event or probability annotation from the corpus described in Section 2.2. This certainly is an idealized dataset containing only sentences with event and/or probability annotation, but we as-

[1] http://nlp.stanford.edu/software/corenlp.shtml

[2] http://metamap.nlm.nih.gov/

sume that we can accurately filter probability sentences. In preliminary experiments we could reliably identify sentences as having a probability or not, using only bag-of-word features (achieving macro-F_1 of 92.3 with a MaxEnt model (Manning and Klein 2003)). The 376 probability sentences are divided into training, development, and test sets (268/52/56 sentences respectively). To establish a baseline, we borrow from the BioNLP/NLPBA 2004 shared task (Kim et al. 2004). Specifically, all events found in the training set are labeled as events in the test set. Then for our initial fully-supervised classification we train a CRF on the labeled training set.

We have chosen CRF++[3] as the CRF implementation for our experiments. We perform feature selection and parameter tuning on the development set. We start with feature selection to determine which context features from positions $t-2$ to $t+2$ are beneficial and settle on the following features: w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}, $\textit{lemma}_{t-1}$, $\textit{lemma}_t$, $\textit{pos}_t$, $\textit{shape}_t$, $\textit{shape}_{t+1}$, plus the CRF++ first-order features (referred to as "bigram" features in CRF++) bi-w_t, bi-$\textit{lemma}_t$, bi-$\textit{pos}_t$. Because our dataset is rather small, we need to restrict our feature set to avoid overfitting (and improve run times). Using these context features and the remaining features from Table 3, we fit the regularization parameter, c, from $\{.5, .3, .1, 1, 3, 5\}$.

4.2 Self-training

Due to the facts that we have a relatively small labeled set and there exists a wealth of unlabeled medical data (e.g., abstracts and articles from MEDLINE), we also want to explore semi-supervised approaches which can leverage the vast amount of data available, and still be directed in a way as to improve classification performance. We include some initial experiments in this direction by employing a standard semi-supervised learning approach, self-training. Self-training is an iterative process that essentially labels instances from an unlabeled set and adds them to the labeled set for the next iteration of model training. The risk of this approach is that if incorrectly labeled instances are included in the labeled set, the self-training will start going in the wrong direction.

We start with a self-training algorithm similar to the one described in (Abney 2007) (see Algorithm 1). The `label` function is our CRF classification. Abney covers variations of this basic self-training algorithm and covers a number of the possible selection criteria. In this paper we try two different heuristic thresholds in the `select` function, which chooses the instances to add to the labeled set, L. To start, our unlabeled set is a sample of 5000 abstracts taken from PubMed.

The first threshold (hereafter "self-training") simply checks if the unlabeled instance (i.e., sentence) has been assigned an event A or B and has probability output from the CRF greater than 0.7, i.e., that $P(\mathbf{y}|\mathbf{x}) > 0.7$. The second threshold (hereafter "PubMed self-training") relies on PubMed. Our intuition is that we can use PubMed to provide additional support in deciding whether events A and B are likely to occur together in a probability statement. For this threshold, first the unlabeled instance must be assigned both events A and B and have a probability greater than 0.3 ($P(\mathbf{y}|\mathbf{x}) > 0.3$). Then we submit two queries to PubMed: 1) the text in event B (e.g., "anti-inflammatory drugs"), and 2) events A and B combined with the AND search operator (e.g., "anti-inflammatory drugs AND cancer"). We use the number of hits returned from each of these queries to estimate how likely it is that event A occurs given event B, $\hat{P}(A|B) = \frac{\text{PubMed hits for } A+B}{\text{PubMed hits for } B}$. If this estimate, $\hat{P}(A|B)$, is greater than 0.1, we add the instance to the labeled set, otherwise it remains in the unlabeled set. These are just a couple of a number of possible selection criteria and we aim to explore better selection criteria in future work, along with more sophisticated semi-supervised learning algorithms.

Algorithm 1 Self-training

L_0 is labeled data, U_0 is unlabeled data
$\mu \leftarrow \texttt{train}(L_0)$
$i \leftarrow 1$
repeat
  $L_i \leftarrow L_{i-1} + \texttt{select}(\texttt{label}(U_{i-1}, \mu))$
  $U_i \leftarrow U_{i-1} - \texttt{select}(\texttt{label}(U_{i-1}, \mu))$
  $\mu \leftarrow \texttt{train}(L_i)$
  $i \leftarrow i + 1$
until stopping criterion is met
return μ

4.3 Development Results

Results are presented in Tables 4 and 5. They are calculated using the `conlleval` Perl script[4] and we report the precision, recall, and F_1 score for individual tokens (Table 4) and for phrases (Table 5). As mentioned previously, the exact boundary is not as critical for our task as in other contexts as the phrases "development of breast cancer", "breast cancer", and "breast cancer risk" should all be normalized to the same node in our BBN model. Therefore, token-based performance results are informative as well. However, it is standard practice to measure entities or text chunks by F_1, despite its drawbacks,[5] which is why we include those results. As can be seen in Table 4, the baseline that we have chosen has good precision and weaker recall. It takes events from the training data that are likely to also be events in the development set. Naturally, events not seen in training will not be labeled, lowering recall.

The CRF method outperforms the baseline with higher overall F_1. Considering the token-based results in Table 4, recall for both events A and B exceeds the baseline, while CRF precision only improves for event A. We test statistical significance of overall F_1, and F_1 for each event, using approximate randomization (Noreen 1989). F_1 differences overall and for event A are significant, but there is no significant difference for event B. The overall results are much

[3] http://crfpp.googlecode.com/svn/trunk/doc/index.html

[4] http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt

[5] http://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html

	Baseline			CRF		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1
Event A	51.9	30.2	38.1	53.4	46.3	49.6
Event B	51.7	25.0	33.7	37.5	32.5	34.8
Prob. term	41.2	35.8	38.3	93.4	92.7	93.0
Overall	46.9	30.5	37.0	64.0	58.3	61.0

	Self-training			PubMed self-training		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1
Event A	51.6	46.3	48.8	50.9	43.4	46.8
Event B	40.0	33.3	36.4	46.2	30.0	36.4
Prob. term	92.0	92.7	92.4	94.7	90.5	92.5
Overall	63.9	58.5	61.1	67.4	55.7	61.0

Table 4: Token-based results on development set

	Baseline			CRF		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1
Event A	23.2	22.0	22.6	58.3	35.6	44.2
Event B	41.8	36.5	39.0	42.5	27.0	33.0
Prob. term	39.4	34.2	36.6	91.4	90.2	90.8
Overall	35.2	31.4	33.2	71.3	54.9	62.1

	Self-training			PubMed self-training		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1
Event A	48.8	35.6	41.2	53.9	35.6	42.9
Event B	47.2	27.0	34.3	41.0	25.4	31.4
Prob. term	91.4	90.2	90.8	91.3	89.0	90.1
Overall	70.0	54.9	61.5	69.6	53.9	60.8

Table 5: Phrase-based results on development set

higher than the baseline due to the performance of probability term extraction, but we are more interested in the performance for events A and B.[6]

The baseline approach tends to find more, but shorter phrases than the CRF: the baseline has 56 event A phrases with an average length of 1.41 words and 55 event B phrases with an average length of 1.05 words, while the CRF has 32 event A phrases with an average length of 2.53 words and 38 event B phrases with an average length of 2.29 words. Because of this, the event B phrase baseline performs quite well. Most of the event B *true positives* are single words; 12 of the 23 true positives come from the phrase "women" which frequently occurs as an event B. If the event B labeled by the CRF is "British women" this counts as a *false positive* and *false negative* with no partial credit for identifying "women".

We ran self-training and PubMed self-training for three iterations because scores began to drop on the third iteration. We report the results after the second iteration, which are the best for both self-training and PubMed self-training. The overall token-based F_1 results for self-training are slightly better than fully-supervised CRF, and the PubMed self-training results do just as well as fully-supervised CRF. These results are modest, but promising. Self-training runs the risk of decreasing performance and has led to negative results in other domains, so these improvements are very encouraging and will push us to improve our selection criteria to get further improvements.

4.4 Error Analysis

We initially argued for measuring F_1 on tokens (Table 4) along with phrases (Table 5) because exact boundaries are less critical for our task. However, looking more closely at the phrase results, the CRF performs well in terms of identifying the correct boundaries. Only 12 of the 204 phrases in the development set had incorrectly labeled boundaries, and of those, four also had erroneous event labels (e.g., A instead of B or vice versa). A few of these errors involve tricky cases with adjectives that are not consistently annotated as part of an event. An example error that illustrates this is the following: (i) "... being $\overbrace{\text{current drinkers}}^{A}$ ..." (from the training corpus) and (ii) "of $\underbrace{\text{current }\overbrace{\text{HER2 testing}}^{A}}_{A}$ may ..." (gold labels above with superscript and classifier labels below with subscript; no annotation means it has the label "O"). In (i), "current" is part of event A, while in (ii) it is not. There are similar errors for other adjectives and also past tense verbs that can have the tags VBD or VBN. For example, "improved" can be simple past (VBD) or a past participle (VBN); in the latter case it often acts as an adjective.

Although these boundary errors doubly penalize F_1 – counting as both a false positive and false negative – they occur relatively infrequently for our task, and will likely have little negative impact in our overall system, which would cluster "current drinkers" and "drinkers" when necessary. There were only three phrases which mislabeled A instead of B and vice versa, and all three contribute to the above errors in boundary detection.

Overall, the majority of the errors involve missing entire phrases. There are 75 event A and B phrases that are not tagged at all (i.e., have label "O") and only 29 phrases which should not be tagged, but have some label A or B. These are the errors we plan to tackle in future work, in particular the high number of false negatives. One solution for this might be to use another classifier with higher recall and stack our CRF model on top using the high-recall classifier output as a new feature. Another option would be to constrain probability sentences to have variables A and B since most errors come from the lack of event A and B labels in the classification and not the misclassification of these events.

4.5 Test Results

We see similar results for the test data. Scores from CRF classification are actually higher for the test set indicating that we did not tune the algorithm to perform well only

[6] We also tested some regular expressions to capture the probability term but the F_1 results were similar to those from this baseline, so we use only this baseline for simplicity.

	Baseline			CRF		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1
Event A	49.5	27.8	35.6	69.6	47.3	56.3
Event B	60.6	27.2	37.6	60.6	38.0	46.7
Prob. term	38.3	33.1	35.5	100.0	84.5	91.6
Overall	47.3	29.3	36.2	78.2	55.8	65.1

	Self-training			PubMed self-training		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1
Event A	64.8	46.8	54.3	65.2	44.4	52.8
Event B	63.5	38.6	48.0	58.4	41.8	48.7
Prob. term	100.0	85.1	92.0	99.2	85.1	91.6
Overall	77.3	56.0	65.0	75.2	56.2	64.3

Table 6: Token-based results on test set

	Baseline			CRF		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1
Event A	22.6	21.9	22.2	53.9	32.8	40.8
Event B	27.0	21.8	24.1	61.5	30.8	41.0
Prob. term	41.1	30.0	34.7	100.0	78.0	87.6
Overall	30.8	25.2	27.7	78.9	50.8	61.8

	Self-training			PubMed self-training		
	Pre.	Rec.	F_1	Pre.	Rec.	F_1
Event A	52.4	34.4	41.5	53.5	35.9	43.0
Event B	61.0	32.1	42.0	60.0	34.6	43.9
Prob. term	100.0	78.0	87.6	100.0	77.0	87.0
Overall	77.6	51.7	62.0	77.0	52.5	62.4

Table 7: Phrase-based results on test set

on the development data. Self-training and PubMed self-training were run for two iterations like for the development set. CRF, self-training, and PubMed self-training all significantly improve over the baseline for overall F_1, as well as for event A and B (using approximate randomization). These three do not significantly differ from each other. Like in the development set, F_1 increases for event B and drops for event A.

5 Related Work

BioNLP event extraction (Kim et al. 2009; Nédellec et al. 2013) is similar to our probability extraction in that it extracts entities, analogous to our probabilistic events A and B, and the relations between entities, i.e., bio-molecular events. Instead, we extract conditional probability relations that include a probability term (see Section 2). In the most recent BioNLP shared task, Björne and Salakoski (2013) achieved the best results for 6 of the 10 tasks, including event extraction. They use a chain of SVM classifiers to detect, e.g., entities, relations, speculation, and negation. We use CRFs to globally optimize labels in the sentence and avoid cascading errors using a chain of classifiers, however, we have not yet been able to try our approach on the GENIA event extraction corpus.

Fiszman et al. (2007) are also interested in extracting risk factors and diseases[7] from medical literature, but differ from us in their approach. They convert the biomedical text into a semantic representation, using the UMLS Semantic Network, and define a set of rules which identify risks and disorders based on the semantic expression. Their approach requires considerable effort and expertise in defining and crafting the semantic rules, however, it could be beneficial to bootstrap our statistical approach using their risks and disorders. In fact, Abacha and Zweigenbaum (2011) conclude that such a combination of statistical approaches with medical domain knowledge sources leads to better results. Their approach is similar to ours, using a CRF with medical features from MetaMap, but the task differs.

There has been previous work that builds Bayesian networks from medical data, however, none extracts conditional probability statements for building BBNs as we do. Sanchez-Graillet and Poesio (2004) build Bayesian networks by focusing on causal relations. Their approach relies on a fixed list of causal relation patterns (from Girju and Moldovan (2002)) that they use for extracting (*cause*,*connective*,*effect*) tuples. The conditional probabilities are estimated using maximum likelihood estimation (MLE) on these tuples and not on the whole dataset. Similarly, Theobald, Shah, and Shrager (2009) use MLE for estimating probabilities but use co-occurrence information from a much larger dataset. This approach for extracting conditional probabilities from the medical text is simple, requiring no real linguistic processing. They simply count the co-occurrences of treatments, diseases, and genes in PubMed data for constructing conditional statements of the form $P(\text{treatment}|\text{disease}, \text{gene})$. In other words, both approaches assume that probabilities are related to co-occurrences in some specific patterns. While practical, and possibly leading to meaningful results, it is slightly troubling to think that, for instance, the probability of developing breast cancer given a person is a man[8] is estimated by counting the ratio of times both terms are present in the same sentence of a chosen set of texts. Corpus selection thus would play a large role in probability estimation. In addition, this would mean that when an author in a medical paper seeks to clarify a point by rephrasing a sentence, he may influence the probability estimation. As presented in this paper, we have chosen a different approach where we seek to extract only explicit probability statements.

6 Conclusions

In this paper, we describe a model for extracting conditional probability statements, in the form of $P(A|B) = x$ from medical texts. This is a difficult task due to the wide variety of forms a conditioned event A and conditioning event B can take. Nonetheless, a satisfactory solution to this problem

[7] *Risks* and *disorders* are subsets of events A and B, but do make up a significant part of each.

[8] Men can develop breast cancer though incidence is lower than for women.

would prove useful, not only to generate input to simplify the process of building Bayesian belief networks for health risk analysis, but also to be able to automatically extract and synthesize probabilistic statements made in a given corpus. The automation of the process makes it possible to perform the analysis over time and identify trends. In addition, other domains, such as finance, auditing, and maintenance, could find time-saving benefits of automatic probability extraction.

Our proposed approach, based on CRFs, improves over an established baseline and shows that this task is better handled as a sequence tagging problem. Using standard sequence tagging features we are able to identify events A and B as well as probability terms. We also got small improvements from using self-training, which we intend to build upon with better selection criteria.

There are a number of other improvements that can be made to reliably use the extracted probability terms and risk events to generate a BBN for a given disease. We would like to look at other semi-supervised approaches to access larger amounts of unlabeled medical data. We also plan to test integer linear programming (ILP) for adding additional constraints to our models (Roth and Yih 2007). Finally, we will explore whether hybrid methods, i.e., those involving both expert knowledge and NLP algorithms can provide further improvements.

Acknowledgments

This project was supported in part through a grant from the Irish Innovation Development Agency (reference 133954, September 2009).

References

Abacha, A. B., and Zweigenbaum, P. 2011. Medical entity recognition: A comparison of semantic and statistical methods. In *Proceedings of BioNLP 2011 Workshop*, 56–64.

Abney, S. 2007. *Semisupervised Learning for Computational Linguistics*. Chapman and Hall/CRC Press.

Aronson, A. R. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In *Proceedings of AMIA, Annual Symposium*, 17–21.

Bikel, D. M.; Miller, S.; Schwartz, R.; and Weischedel, R. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 194–201. Washington, DC, USA: Association for Computational Linguistics.

Björne, J., and Salakoski, T. 2013. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 shared task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, 16–25. Sofia, Bulgaria: Association for Computational Linguistics.

Cooke, R., and Goossens, L. 1999. Procedures guide for structured expert judgment. Technical Report EUR18820, Brussels-Luxembourg: Commission of the European Communities.

Deleris, L. A.; Yeo, G. L.; Seiver, A.; and Paté-Cornell, M. E. 2006. Engineering risk analysis of a hospital oxygen supply system. *Medical Decision Making* 26(2):162–172.

Fiszman, M.; Rosemblat, G.; Ahlers, C. B.; and Rindflesch, T. C. 2007. Identifying risk factors for metabolic syndrome in biomedical text. In *Proceedings of the AMIA Annual Symposium*, 249–253.

Fuller, G. 2005. Simulconsult: www.simulconsult.com. *Journal of Neurology, Neurosurgery & Psychiatry* 76(10):1439–1439.

Girju, R., and Moldovan, D. I. 2002. Text mining for causal relations. In *FLAIRS Conference*, 360–364.

Hoffer, E. P.; Feldman, M. J.; Kim, R. J.; Famiglietti, K. T.; and Barnett, G. O. 2005. DXplain: patterns of use of a mature expert system. *AMIA Annual Symposium Proceedings* 2005:321–324. PMID: 16779054 PMCID: PMC1560464.

Kim, J.-D.; Ohta, T.; Tateisi, Y.; and Tsujii, J. 2003. GENIA corpus–a semantically annotated corpus for bio-textmining. *Bioinformatics* 19(suppl 1):i180–i182.

Kim, J.-D.; Ohta, T.; Tsuruoka, Y.; Tateisi, Y.; and Collier, N. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, JNLPBA '04, 70–75. Stroudsburg, PA, USA: Association for Computational Linguistics.

Kim, J.-D.; Ohta, T.; Pyysalo, S.; Kano, Y.; and Tsujii, J. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, 1–9. Boulder, Colorado: Association for Computational Linguistics.

Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 282–289. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Manning, C. D., and Klein, D. 2003. Optimization, maxent models, and conditional estimation without magic. In *Proceedings of HLT-NAACL*, 8.

Nédellec, C.; Bossy, R.; Kim, J.-D.; Kim, J.-J.; Ohta, T.; Pyysalo, S.; and Zweigenbaum, P. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, 1–7. Sofia, Bulgaria: Association for Computational Linguistics.

Noreen, E. 1989. *Computer-intensive methods for testing hypotheses: An introduction*. Wiley.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pietzsch, J. B., and Pat-Cornell, M. E. 2008. Early technology assessment of new medical devices. *International Journal of Technology Assessment in Health Care* 24:36–44.

Roth, D., and Yih, W. 2007. Global inference for entity and relation identification via a linear programming formulation. In Getoor, L., and Taskar, B., eds., *Introduction to Statistical Relational Learning*. MIT Press.

Sanchez-Graillet, O., and Poesio, M. 2004. Acquiring Bayesian networks from text. In *LREC*.

Settles, B. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications*, JNLPBA '04, 104–107. Stroudsburg, PA, USA: Association for Computational Linguistics.

Sha, F., and Pereira, F. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, 134–141. Stroudsburg, PA, USA: Association for Computational Linguistics.

Steele, S.; Bilchik, A.; Eberhardt, J.; Kalina, P.; Nissan, A.; Johnson, E.; Avital, I.; and Stojadinovic, A. 2012. Using machine-learned Bayesian belief networks to predict perioperative risk of clostridium difficile infection following colon surgery. *Interact J Med Res* 1(2):e6.

Theobald, M.; Shah, N.; and Shrager, J. 2009. Extraction of conditional probabilities of the relationships between drugs, diseases, and genes from PubMed guided by relationships in PharmGKB. In *2009 AMIA Summit on Translational Bioinformatics*, 124–128. Grand Hyatt, San Francisco: American Medical Informatics Association.

van der Gaag, L. C.; Renooij, S.; Witteman, C.; Aleman, B. M.; and Taal, B. G. 2002. Probabilities for a probabilistic network: a case study in oesophageal cancer. *Artificial Intelligence in medicine* 25(2):123–148.

Wang, C., and Fan, J. 2014. Medical relation extraction with manifold models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 828–838. Baltimore, Maryland: Association for Computational Linguistics.

Warner, H.; Haug, P.; Bouhaddou, O.; Lincoln, M.; Sorenson, D.; Williamson, J.; and Fan, C. 1988. Iliad as an expert consultant to teach differential diagnosis. In *Proceedings of the 12th Symposium on Computer Applications in Medical Care (SCAMC), IEEE Computer Society Press*, 371–376.

Watt, E. W., and Bui, A. A. 2008. Evaluation of a dynamic Bayesian belief network to predict osteoarthritic knee pain using data from the osteoarthritis initiative. *AMIA Annu Symp Proc* 788–792.