Semi-Automatic Construction of Thyroid Cancer Intervention Corpus from Biomedical Abstracts

Wuthipong Kongburan, Praisan Padungweang, Worarat Krathu, Jonathan H. Chan

School of Information Technology
King Mongkut's University of Technology Thonburi
Agenda

INTRODUCTION

PROPOSED METHOD

RESULT AND DISCUSSION

CONCLUSION AND FUTURE WORK
Agenda

- INTRODUCTION
  - Overview of text mining and NER
  - Thyroid Cancer and Medical Intervention
  - Rationale and Motivation
  - Research questions

PROPOSED METHOD
RESULT AND DISCUSSION
CONCLUSION AND FUTURE WORK
Thyroid cancer

The cells of the thyroid gland become abnormal, grow uncontrollably, and form a mass of cells called a tumor.

Thyroid neoplasms represent almost 95% of all endocrine tumors.

Women are three times more likely than men.

In Bangkok, Thailand, the number of new cases of thyroid cancer is 6 per 100,000 men and women per year.

It is now ranked as the 8th most common cancer in women in the United States and 4% increase annually.
Medical Intervention

- “Any objects or acts of **treatment** and **preventive care** of intervening, interfering or interceding with the intent of modifying the outcome”
- Such typical treatment - the side effects.
- Medical interventions
  - alternative treatment
  - increasing patient satisfaction
  - lead to better health outcomes
Thyroid cancer publications in PubMed

![Graph showing the increase in publications over years](image-url)
Overview of Biomedical Text mining

Literature sources:
- MEDLINE
- ScienceDirect
- MedWatch
- ...

Information Retrieval

Named Entity Recognition

Relevant concepts:
- PDK4
- carcinoma
- breast cancer
- glycolysis
- BRAF
- ...

Information Extraction

Knowledge Discovery

Knowledge:
- breast cancer
- glycolysis
- PDK4
- BRAF

---

known

hypothesis
Named Entity Recognition (NER)

- Classify tokens into predefined categories such as person name, organizations, diseases, proteins.
Rationale and Motivation

Training data → ML-based Text mining → Classifier/Model
Research questions

- How can we construct a thyroid cancer intervention corpus in reasonable time?

- What are the criteria to measure the performance of corpus?
Agenda

INTRODUCTION

- PROPOSED METHOD
  - A construction workflow of corpus
  - Basic definitions of entities
  - Evaluation of corpora
    - 10-fold cross validation
    - Leave-one(intervention)-out cross validation

RESULT AND DISCUSSION

CONCLUSION AND FUTURE WORK
A construction workflow of thyroid cancer intervention corpus

PubMed query:
```
(((thyroid neoplasms) OR thyroid cancer) OR thyroid carcinoma) OR thyroid malignancies) OR papillary) OR follicular) OR medullary) OR anaplastic) AND treatment) AND hasabstract[text] AND Humans[Mesh] AND English[lang])"
```

The query limits the results to abstracts of human studies in English.

Sorted by relevance
A construction workflow of thyroid cancer intervention corpus

The initial number of 50 abstracts is arbitrary, subject to the limitation of human annotator that can be done in reasonable time.

To determine entity boundaries in a text by sentence splitting and define the data in token form.
A construction workflow of thyroid cancer intervention corpus

Each token or word is tagged with three classes of biomedical entities: INTERVENTION, DISEASE, and O.

The annotation process is carried out following the annotation guideline.
Basic definitions of entities

- **INTERVENTION** means a treatment and preventive care object or action that performed by doctor or other clinician targeted at a thyroid cancer patient.
  - Non-drug therapy: e.g., surgery, radioactive iodine treatment, external beam radiation therapy, and
  - Drug therapy: e.g., anti-angiogenesis drugs, target inhibitors, chemotherapy, and thyroid hormone replacement therapy.

- **DISEASE** means a disorder in which the cells of the thyroid gland become abnormal, grow uncontrollably, and form a mass of cells called a tumor. Including
  - Common name and tumors name of thyroid cancer: e.g., thyroid cancer, thyroid malignancies, thyroid carcinoma, thyroid tumors.
  - Group names of thyroid cancer: e.g., differentiated thyroid carcinoma, undifferentiated thyroid cancer.
  - Sub types or other forms of thyroid cancer: e.g., follicular cancer, papillary thyroid cancer, Hurthle cell carcinoma, medullary carcinoma, anaplastic carcinoma, insular thyroid carcinoma, thyroid lymphoma, and thyroid sarcoma.

- **Other** terms which not refer to **INTERVENTION** and **DISEASE** will be tagged as ‘O’.

A construction workflow of thyroid cancer intervention corpus

Stanford NER provides a general implementation of linear chain Conditional Random Fields (CRFs) sequence models.
A construction workflow of thyroid cancer intervention corpus

The number of n equals to 30 for each experiment round which is the arbitrarily selected in our current experiment.
A construction workflow of thyroid cancer intervention corpus
A construction workflow of thyroid cancer intervention corpus

To filter only abstracts that contain new *INTERVENTION* out of the manually annotated datasets.
A construction workflow of thyroid cancer intervention corpus

The automatic annotated abstracts will be judged by annotators together with the annotation guideline.

This step can reduce the effort of manually annotated.
A construction workflow of thyroid cancer intervention corpus

The result of each round of incremental model is a corpus. It is a combination of manually annotated datasets and automatic tagging with manual correction datasets.
A construction workflow of thyroid cancer intervention corpus

The process will be repeated again with next 30-first abstracts, until the performance of a new corpus and current corpus have no significant difference.
Incremental construction

<table>
<thead>
<tr>
<th>Round</th>
<th>Current corpus size</th>
<th>#Tokens unique interventions</th>
<th>New n-first abstracts</th>
<th>#Abstracts contain new intervention</th>
<th>#Abstracts after manual correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>312</td>
<td>1-30</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>59</td>
<td>339</td>
<td>31-60</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>67</td>
<td>356</td>
<td>61-90</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>73</td>
<td>379</td>
<td>91-120</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>81</td>
<td>400</td>
<td>121-150</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>89</td>
<td>413</td>
<td>151-180</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>99</td>
<td>460</td>
<td>181-210</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>112</td>
<td>501</td>
<td>211-240</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>9</td>
<td>125</td>
<td>525</td>
<td>241-270</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>135</td>
<td>546</td>
<td>271-300</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>143</td>
<td>564</td>
<td>301-330</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>153</td>
<td>594</td>
<td>331-360</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>162</td>
<td>619</td>
<td>361-390</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>167</td>
<td>632</td>
<td>391-420</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>15</td>
<td>174</td>
<td>648</td>
<td>421-450</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>16</td>
<td>181</td>
<td>655</td>
<td>451-480</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>17</td>
<td>188</td>
<td>675</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Research questions

- How can we construct a thyroid cancer intervention corpus in reasonable time?

- What are the criteria to measure the performance of corpus?
INTRODUCTION

PROPOSED METHOD
- A construction workflow of corpus
- Basic definitions of entities
- Evaluation of corpora
  - 10-fold cross validation
  - Leave-one(intervention)-out cross validation

RESULT AND DISCUSSION

CONCLUSION AND FUTURE WORK
Performance Evaluation
10-fold cross validation

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>Actual class</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERVENTION</td>
<td>DISEASE or O</td>
</tr>
<tr>
<td>True Positive(TP)</td>
<td>False Positive(FP)</td>
</tr>
<tr>
<td>False Negative(FN)</td>
<td>True Negative(TN)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>Actual class</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISEASE</td>
<td>INTERVENTION or O</td>
</tr>
<tr>
<td>True Positive(TP)</td>
<td>False Positive(FP)</td>
</tr>
<tr>
<td>False Negative(FN)</td>
<td>True Negative(TN)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>Actual class</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>INTERVENTION or DISEASE</td>
</tr>
<tr>
<td>True Positive(TP)</td>
<td>False Positive(FP)</td>
</tr>
<tr>
<td>False Negative(FN)</td>
<td>True Negative(TN)</td>
</tr>
</tbody>
</table>

Precision (P) = \( \frac{TP}{TP + FP} \), Recall (R) = \( \frac{TP}{TP + FN} \), F1-score (F) = 2 \( \frac{precision \times recall}{precision + recall} \)
10-fold cross validation

Sequentially select

Randomly select
INTRODUCTION

PROPOSED METHOD

- RESULT AND DISCUSSION
  - How many abstracts are suitable for constructing the corpus?

CONCLUSION AND FUTURE WORK
The average performance of three entity classes by using 10-fold cross validation.
The change rate of average F1-score of three entity classes
The average performance of 143 model

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERVENTION</td>
<td>0.923</td>
<td>0.666</td>
<td>0.768</td>
</tr>
<tr>
<td>DISEASE</td>
<td>0.982</td>
<td>0.915</td>
<td>0.946</td>
</tr>
<tr>
<td>O</td>
<td>0.974</td>
<td>0.996</td>
<td>0.984</td>
</tr>
<tr>
<td>Average</td>
<td>0.959</td>
<td>0.859</td>
<td>0.906</td>
</tr>
</tbody>
</table>

in a case of 143 abstracts there are 41,657 tokens consist of
- 2,473 INTERVENTION (564 unique intervention tokens),
- 1,996 DISEASE, (94 unique tokens of disease), and
- 37,188 O.
Leave-one(intervention)-out cross validation

All leave out keywords are the words that describe common interventions applied for thyroid cancer e.g. chemotherapy, radioiodine, radioactive, radiotherapy, drug, vandetanib, cabozantanib, sorafenib, inhibitors, surgery, ablation, resection, dissection, and thyroidectomy.
The accuracy of discovery new interventions by using leave-one-out cross validation
Result and discussion

- Based on the experimental results, the corpus with 143 abstracts is a suitable size of corpus for
  - Acceptable model in performance metrics.
  - Discovery new interventions (overfitting avoidance)
Confusion matrix of ‘radioactive’ keyword in leave-one-out cross validation using 143 model

<table>
<thead>
<tr>
<th>Actual class</th>
<th>Predicted class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>8668</td>
</tr>
<tr>
<td>INTERVENTION</td>
<td>191</td>
</tr>
<tr>
<td>DISEASE</td>
<td>19</td>
</tr>
</tbody>
</table>

False negatives and false positive mainly fall into
- mislabeled (e.g., ‘*total lobectomy*’ is annotated as O),
- lacking in entity’s boundaries (e.g., ‘*central neck node dissection*’ only *dissection* annotated as *INTERVENTION*), and
- case sensitive (e.g., ‘*Surgery*’ is annotated as O instead *INTERVENTION*).
Agenda

INTRODUCTION

PROPOSED METHOD

RESULT AND DISCUSSION

- CONCLUSION AND FUTURE WORK
Result and discussion

- The production of a gold standard corpus requires a significant amount of manual curation work

**Contribution**
- a semi-automatic semantic annotation approach to construct a thyroid cancer intervention corpus together with a criteria of identifying reasonable performance

**Future work**
- Improve performance of the corpus for example, diversify the keywords search, remove sentences without annotations, and use different number of abstracts in the experiment
- Apply our approach to extract relationships between general biomedical entities
THANK YOU FOR YOUR ATTENTION ANY QUESTIONS