Automatical Classification of Pathological Reports

Janneke van der Zwaan, Erik Tjong Kim Sang and Maarten de Rijke
University of Amsterdam

Questions from a guest @ UvA

We have a huge collection of abstracts of medical reports which were enriched with keywords. We want to make the contents of the abstracts as accessible as possible.

- How can we automatically generate keywords from abstracts?
- With respect to searching the abstracts: should we continue focusing on assigning keywords or switch to full text search?

PALGA

- PALGA: Dutch national network and registry of histopathology and cytopathology (studies of tissue and cell samples)
- Pathologists create reports of every pathological test done in The Netherlands
- Reports have been stored in a national database since 1971

The PALGA database

About one million report abstracts per year about tissue and cell tests performed by pathologists in The Netherlands:

- report conclusion (abstract)
- diagnosis line (keywords)
- other information: document id, patient id, date

Example from the database

ID	393191785
Patient ID	PATIENT-23-m
Date	07 - 1990
Abstract	Huidexcisie para-orbitaal links: basaalcelcarcinoom van het solide type. Tumorcellen reiken tot in de excisieranden.
Keywords	huid * gelaat * links * excisie * basaalcelcarcinoom

PALGA coding system

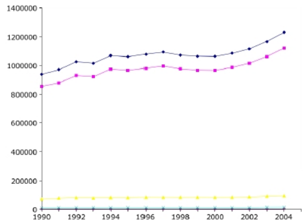
The keywords contain three compulsory fields:

- topography: sample origin (location in body)
- procedure: method used for obtaining the sample
- diagnosis

The fields can contain more than one keyword.

There is a thesaurus which restricts the keywords that can be used.

Number of records per year



Goal

To predict the keywords for an abstract as accurately as possible.

Approach

- We use support vector machines with linear kernels.
- Data was restricted to colon data with “clean” fields (about 0.5 million records)
- Random training/test data split in 75%/25%
- Predict occurrences of the 132 most frequent terms
- Use different data representations

Evaluation

- We compute precision and recall rates for term predictions
- Predict terms are correct when they match corpus terms

Experiment variants

- Baseline: represent abstracts as *bag of words*
- Alternatives: tf-idf, thesaurus terms, terms+parents, word bigrams and combinations of these
- We also tested the effects of stemming and compound splitting
- Preprocessing with rule-based term identification
- Feed back predicted terms to the learner
- Change parameters of the learner (kernel type)

Experiment results

	Precision	Recall	F-measure
bag of words (bow)	83.28%	72.92%	77.76%
tf-idf	83.14%	73.71%	78.14%
tf-idf+terms+parents	83.35%	74.09%	78.45%
bigrams+terms	84.85%	75.56%	79.94%
bow+polynomial kernel	84.89%	76.11%	80.26%

Polynomial kernels outperform linear kernels but learning polynomial models requires a lot of time.

Discussion

The recall scores are considerably lower than the precision scores.

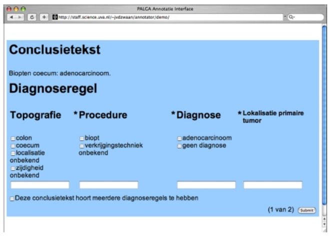
This could be caused by incomplete abstracts. Keywords assignment is based on the complete reports (which we do not have access to). If the abstract is incomplete, the correct keywords cannot be determined.

We decided to test this by making expert pathologists assign keywords to abstracts.

Comparison with domain experts

- Two expert pathologists labeled 1000 abstracts each
- Each took 4.5 hours for the task (16 seconds per abstract)

Web interface for tagging by experts


PALCA Annotatie Interface
http://staff.science.uva.nl/~jvdzwaan/annotator/demo/

Conclusietekst

Biopten coecum: adenocarcinoom.

Diagnoseregel

Topografie
□ colon
□ coecum
□ localisatie onbekend
□ zijdigheid onbekend

*Procedure
□ biopt
□ verkrijgingstechniek onbekend

*Diagnose
□ adenocarcinoom
□ geen diagnose

*Lokalisatie primaire tumor

□ Deze conclusietekst hoort meerdere diagnoseregels te hebben

(1 van 2) Submit

Expert results

A comparison of the keywords assigned by the two domain experts with the original keywords in the PALGA database:

	Precision	Recall	F-measure
bag of words	83.62%	72.87%	77.87%
bigrams+terms	84.88%	75.47%	79.90%
Pathologist A	71.75%	72.54%	72.14%
Pathologist B	66.75%	67.33%	67.04%

Result discussion

- As expected: recall scores for experts are low
- But precision scores are also low!
- Experts often assign related terms rather than the correct ones
- Experts more often agree with each other than with the corpus

Concluding remarks

- SVMs assign keywords to medical texts with P=85 and R=75
- Baseline performance was P=84 and R=73: hard to improve
- Human experts obtain P=70 and R=70
- Choosing appropriate keywords for text is not an easy task!

THE END