A Section Title Authoring Tool for Clinical Guidelines

Mark Truran
School of Computing
Teesside University
United Kingdom
m.a.truran@tees.ac.uk

Gersende Georg
Haute Autorité de Santé
Saint-Denis La Plaine Cedex
France
g.georg@has-sante.fr

Marc Cavazza
School of Computing
Teesside University
United Kingdom
m.cavazza@tees.ac.uk

Dong Zhou
Hunan University of Science
and Technology
China
dongzhou1979@hotmail.com

ABSTRACT

Professional users of medical information often report difficulties when attempting to locate specific information in lengthy documents. Sometimes these difficulties can be attributed to poorly specified section titles which fail to advertise relevant content. In this paper we describe preliminary work on a software plug-in for a document engineering environment that will assist authors when they formulate section-level headings. We describe two different algorithms which can be used to generate section titles. We compare the performance of these algorithms and correlate our experimental results with an evaluation of title quality performed by domain experts.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Medical information systems; I.7.2 [**Document and Text Processing**]: Document Preparation

General Terms

Documentation, Measurement, Human Factors

Keywords

Clinical guideline, section, title, content, quality

1. INTRODUCTION

Clinical guidelines are expert-level documents which describe best practices for the diagnosis and treatment of specific conditions. They are produced by groups of experts under the auspices of health regulatory bodies. Accessing the contents of these documents can be a challenging task for health professionals, who often have to hunt for specific information within a lengthy guideline. As a result, online access to a complete guideline in PDF format frequently proves unproductive [8]. Furthermore, due to the regulatory nature of the information concerned, arbitrary content summarisation (as a possible solution to the document access problem) is simply not feasible. There is a genuine risk that important medical information relevant to patient health and safety may be omitted from auto-generated summaries.

Within HAS (the French National Authority for Health), guidelines are authored using G-DEE, a purpose built document engineering environment [4]. Recently, HAS has begun supplementing each PDF guideline it releases with a limited depth hypertext containing a subset of the same information (in French: *recommandations cliquables*, or reco2clics[1]). Experience with this new document format has reinforced the *vital* importance of section titles as signposts for readers interested in specific sub-topics within a guideline. Poorly specified section titles confuse or mislead guideline readers, thereby reducing the usefulness of the document as a whole.

In this paper we present preliminary work on an authoring tool for clinical guidelines which addresses the problem of poorly specified section titles. This tool will function in an unobtrusive manner, suggesting possible title words to G-DEE users as they author new guidelines. The novel contributions of this work include (1) a contrastive experiment which exports two techniques popular in article-level title generation to the sectional level and (2) an evaluation of title quality, performed by domain experts, correlated against our experimental results.

2. RELATED WORK

Most recent work on the automatic generation of titles has concentrated on *article-level* title generation. For example, Witbrock and Mittal [10] generated titles for a large collection of news-wire articles using a technique inspired by statistical translation (see also [1]). Processing the text of each article in turn, they built a statistical model describing the relationship between the occurrence of a specific term in the document and the occurrence of the same term in

[1]HAS currently host 12 reco2clics guidelines on their web site, each attracting 150-200 downloads per week. In a recent survey, 68% of respondents agreed that the new document format facilitates document access.

DocEng'12, September 4–7, 2012, Paris, France.

the document title. This allowed the authors to estimate the conditional probability of any given term t appearing in the title by calculating $P(t \text{ in title} \mid t \text{ in document})$. Having trained their model using a corpus of 8000 articles, they generated titles for 1000 unseen documents, achieving a respectable overlap with the 'actual' titles.

In [5], Jin and Hauptmann compared the Naïve Bayesian approach described above with three alternative techniques for generating titles:

1. A model based on concepts drawn from the field of Information Retrieval (IR), which included various term weighting measures (see §3.1).
2. An approach that utilised the K-Nearest Neighbour (KNN) algorithm, as applied to topic classification.
3. An algorithm exploiting Expectation-Maximisation (EM), which was used to build a statistical translation model between the 'concise' language of the title and the 'verbose' language of the document.

In an experiment involving over 50,000 document and title pairs, the IR approach to the title word selection problem was declared the most effective, followed closely by the Naïve Bayesian model and KNN.

In a slightly different context, Chakrabarti et al. used a statistical model that exploited multiple sources of information to create link-titles for result URLs [2]. In addition to the content of the the web page indicated by the URL, the authors utilised del.icio.us tags, anchor tags and queries associated with the web page via click-through logs to generate 'quicklinks'. In an empirical evaluation, their statistical model outperformed various competing approaches, including the technique described by Banko et al. [1] (see above).

2.1 Evaluation of generated titles

Systems that generate title words are often evaluated using the F1 score [7]. Assuming you have a human specified title T_{human} and an auto-generated title T_{auto}, the F1 score for T_{auto} is:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

where *precision* is the number of words in T_{auto} that match words appearing in T_{human} divided by the count of words in T_{auto}, and *recall* is the number of correctly generated words in T_{auto} divided by the number of words in T_{human}. This calculation produces a value in the range of [0-1], a value of 1 being the best score and 0 the worst. Providing a useful baseline for title generation algorithms, Jin and Hauptmann recorded F1 scores of 0.226 using the IR model and 0.201 for the Bayesian model in [5].

3. METHODOLOGY

The data used in this experiment was taken from 28 clinical guidelines published in PDF format by the French Health Authority between January 2008 and December 2011. Each guideline was processed in the following way:

1. All of the text in each guideline was extracted using Apache PDFBOX[2] (a Java class library for manipulating PDF documents).
2. A *stop list*[3] was used to remove words with little informational value e.g. determiners, conjunctions, prepositions and pronouns [9].
3. All inflected words were reduced to their root stem e.g. 'mountaineering' and 'mountaineer' were reduced to 'mountain'. We used a stemming algorithm[4] based on Porter's SNOWBALL project [6].
4. We extracted all of the sections from the document using the section titles as delimiters. We saved each section to a separate file.
5. We removed all of the sections with *generic* titles (e.g. 'Sommaire', 'Recommandations', 'Introduction', 'Appendix') to isolate the *ideational* section headings i.e. titles that indicate the content of the section [3].

At the end of this process we had 435 sections (avg. 15.5 sections per guideline) containing 7429 unique terms.

3.1 Algorithms

We used two different techniques when generating title words. The first technique, which is known as TF*IDF, belongs to the field of information retrieval. It involves calculating the number of times a particular word has appeared in a document (TF, or *term frequency*). This value is balanced against the popularity of the word across all of the documents (IDF, or *inverse document frequency*). As shown in Algorithm 1, we make minor modifications to this statistic to generate title words from document *sections*.

Algorithm 1 Generating title words using TF * IDF

Require: s, a stemmed section with stop words removed
Require: n, the number of title words required
  create empty map m (string $\rightarrow$ integer)
  calculate number of sections in the corpus as c
  for all unique words w in s as $w_1, w_2 \ldots w_n$ **do**
    calculate tf as count of $w \in s$
    count sections in corpus that contain w as *df*
    calculate *idf* as LOG (c / *df*)
    put s, $(idf * tf)$ in m
  end for
  order m by $(idf * tf)$, in descending order
  select n topmost entries in m as title words

So, where a section s has 100 terms in total, and a term x is mentioned 7 times, the term frequency of $x \in s = 0.07$. If that term occurs in 25 out of 90 sections in the corpus, its final weight is LOG(90/25) = 0.55 * 0.07 = 0.038. Note that inverse document frequency diminishes the importance of terms that are common across the corpus.

The second technique we used when generating title words was based on a Naïve Bayesian approach. We followed the approach of Witbrock and Mittal [10], creating a limited vocabulary statistical model that described the relationship between source text units in the section and target text units in the title. Generating title words using this approach involved two distinct stages. In the first stage, we trained the model (see Algorithm 2). In the second stage, we used the model to generate title words (see Algorithm 3). Assume a

[2]http://pdfbox.apache.org/

[3]http://members.unine.ch/jacques.savoy/clef/frenchST.txt

[4]http://morphadorner.northwestern.edu/morphadorner/

term x appears in 9 sections, and 4 of those sections have titles that also contain x. According to our algorithm, the conditional probability $P(x$ in title $\mid x$ in section) would be $= 4/9 = 0.44$.

Algorithm 2 Training the model

Require: c, a corpus of sections
  create empty map m (string $\rightarrow$ integer)
  for all sections s in c as $s_1, s_2 \ldots s_n$ **do**
    for all unique words w in the section $w_1, w_2 \ldots w_n$ **do**
      if $w \notin m$ **then**
        put $(w, 0)$ in m
      end if
      if $w \in$ title of s **then**
        increment w in m
      end if
    end for
  end for

3.2 Training corpus and experimental corpus

In this experiment we separated the data into two parts. The *training corpus* was made up from sections extracted from 27 clinical guidelines (approx. 420 sections). We used this corpus to train the statistical model described in Algorithm 2 and to generate the document frequency statistics needed in Algorithm 1. The *experimental corpus* was made up from sections extracted from just 1 clinical guideline (approx. 15 sections). We used this corpus to test the performance of our title generation algorithms. We iterated through the experimental corpus, generating title words for every section using both algorithms. We evaluated the performance of the two algorithms using the precision measure described in §2.1. We repeated this procedure 28 times, rotating the data so that each guideline in the collection 'took its turn' as the experimental corpus.

Algorithm 3 Generating title words using the model

Require: s, a stemmed section with stop words removed
Require: n, the number of title words required
Require: $alg2$, the map from Algorithm 2
  create empty map m (string $\rightarrow$ integer)
  for all unique words w in s as $w_1, w_2 \ldots w_n$ **do**
    retrieve value for w in $alg2$ as $hits$
    count sections in corpus that contain w as df
    put $(s, hits/df)$ in m
  end for
  order m by $hits/df$, in descending order
  select n topmost entries in m as title words

Here is a worked example that describes our evaluation technique. The original section title for guideline 55 §3.1 is 'Quelles sont les situations pouvant faire l'objet d'une prescription médicamenteuse par téléphone lors de la régulation médicale?', *trans*: 'In which situations can drug prescriptions be made over the telephone by medical staff in charge of emergency helpline and dispatching?' Following the application of a stop list and a stemming algorithm, this is reduced to 'situat pouv fair l'objet d'une prescript médic téléphon lor régul'. We pass the text for this section to the title generation algorithms with the parameter n arbitrarily set to the length of the pre-processed title (i.e., 10 terms). The algorithm exploiting TF*IDF produces 5 matching terms (régul, médic, prescript, situat, téléphon) and 5 non-matching terms, thereby achieving a precision score of 0.5. The NBL algorithm produces two matching terms (régul, médic) and 8 non-matching terms, achieving a lower precision score of 0.2. Note that since n is equal to the length of the pre-processed title, the precision measure is equivalent to both recall and F1.

	TF*IDF	NBL
MINIMUM	0.07	0.0
LOWER QUARTILE	0.10	0.004
MEDIAN	0.20	0.06
UPPER QUARTILE	0.23	0.11
MAXIMUM	0.347	0.306

Table 1: Precision of TF*IDF and NBL algorithms across 28 iterations of experiment

4. RESULTS AND ANALYSIS

The results of the experiment are described in Table 1. As illustrated, the algorithm exploiting TF*IDF produced the best performance, achieving an average precision of 0.18 i.e., it accurately predicted 18% of the title words used by authors. By comparison, title generation based on a Naïve Bayesian model with limited vocabulary (NBL) was inferior, achieving an average precision of 0.07. This was the result we expected. The two techniques used in this experiment, and in particular the technique reliant on a Naïve Bayesian model, generally produce results which are positively related to the size and quality of the training corpus. Our training corpus had less than 500 sections. Researchers developing models to predict the title of articles commonly use tens of thousands of documents [5, 1]. Given the above, we consider these to be satisfactory scores. More importantly, the fundamental aim of the experiment - i.e. proof of concept for section level title generation - was comfortably achieved.

4.1 Expert evaluation

Most studies of this type will score the generated title words against the original title, thereby implicitly assuming that the original title was a 'good' one (see §2.1). We decided to challenge this assumption. We asked a panel of 20 domain experts to evaluate 20 section titles randomly selected from the corpus. The judges scored each section title using a 5-point Likert scale which described the relevance of the title to the section text. The highest ranked section title was an extremely specific heading taken from guideline 55 §5.4 (in French: 'Comment assurer la traçabilité de l'entretien téléphonique?', *trans*: 'How to log and trace calls made to emergency helplines', average human score 3.9/4). The lowest ranked section title, extracted from guideline 78 §5.1 (in French 'Aspects physiques', *trans*: 'Physical aspects', average human score 1.5/4) was perhaps too generic for the panel given the context.

When we correlated the human scores with our experimental results, we noticed a possible *inverse relationship* between human satisfaction and the performance of the TF*IDF algorithm. As described in Table 2 and illustrated in Figure 1, the TF*IDF score for the bottom 10 section titles (as ranked by humans) was 0.31, which is approximately 3 times higher than the same score for the top 10 titles. A

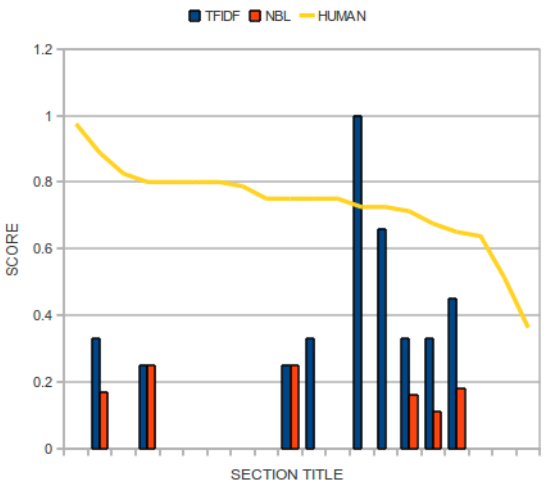

Figure 1: Comparison of title generation performance with human judgements. Human judgements have been transformed from [0,4] to [0,1] range. Zero values indicate that TF*IDF/NBL algorithm matched zero title words.

RANK ASSIGNED BY EXPERTS	1-10	11-20
AVERAGE TF*IDF SCORE	0.083	0.310

Table 2: Correlation of average TF*IDF scores with the rank assigned by the human judges. Titles were sorted in descending order by human rating and split into two equal sized groups, 1-10 and 11-20.

(very tentative) hypothesis given the limited data - algorithms exploiting TF*IDF may have a tendency to produce title words that are unpopular with humans! This is an intriguing finding because a number of popular title generation algorithms rely heavily on TF*IDF (see §2). If this inverse relationship is replicated in a larger study, it may prompt a re-evaluation of these techniques in favour of algorithms exploiting other sources of information (e.g. [2]).

5. FURTHER WORK

Given the understated performance of the Naïve Bayesian model in this experiment, further work should include an attempt to expand the training resources used to generate section-title representations. Possible repositories of useful training data include the French Wikipedia medical portal and CISMeF[5]. Future work could also include the use of a domain specific ontology or French medical thesaurus to enrich the suggestions made by the TF*IDF algorithm.

6. CONCLUSION

In this paper we have described preliminary work on a software plug-in for the G-DEE document engineering environment that will assist authors as they formulate section-level headings. These titles are vitally important to readers as signposts marking the position of specific sub-topics within a clinical guideline. Our work exports two techniques popular in article-level title generation to the sectional level. Our primary findings indicate that an algorithm based on TF*IDF will outperform a Naïve Bayesian model when training resources are meagre. This is not surprising. Our secondary findings, which disclose a possible inverse relationship between human satisfaction and weighted term frequency analysis, are more interesting. An extensive follow up study featuring qualitative survey questions (e.g. what makes this a good/bad section title?) is indicated.

[5]http://www.chu-rouen.fr/cismef/

7. ACKNOWLEDGEMENTS

Many thanks to HAS staff from SBPP/SEESP units who participated in this study. The authors would also like to thank the anonymous reviewers for their valuable comments.

8. REFERENCES

[1] M. Banko, V. O. Mittal, and M. J. Witbrock. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 318–325, Stroudsburg, PA, USA, 2000. ACL.

[2] D. Chakrabarti, R. Kumar, and K. Punera. Generating succinct titles for web urls. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 79–87, New York, NY, USA, 2008. ACM.

[3] S. Gardner and J. Holmes. From section headings to assignment macrostructures in undergraduate student writing. In *Thresholds and Potentialities of Systemic Functional Linguistics: Multilingual, Multimodal and Other Specialised Discourses*, pages 268–290. Edizioni Universita di Trieste, 2010.

[4] G. Georg and M.-C. Jaulent. A document engineering environment for clinical guidelines. In *Proceedings of the 2007 ACM symposium on Document engineering*, DocEng '07, pages 69–78, New York, NY, USA, 2007. ACM.

[5] R. Jin and A. G. Hauptmann. Learning to select good title words: An new approach based on reverse information retrieval. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 242–249, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[6] M. F. Porter. Readings in information retrieval. chapter An algorithm for suffix stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[7] C. J. V. Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[8] A. H. Rø svik and H. P. Fosseng. Usability testing of clinical guidelines. Guidelines International Network Conference 2011, Korea University, Seoul, Korea, 2011.

[9] J. Savoy. A stemming procedure and stopword list for general french corpora. *Journal of the American Society for Information Science*, 50:944–952, 1999.

[10] M. J. Witbrock and V. O. Mittal. Ultra-summarization (poster abstract): a statistical approach to generating highly condensed non-extractive summaries. In *Proceedings of the 22nd annual international ACM SIGIR conference*, SIGIR '99, pages 315–316, New York, NY, USA, 1999. ACM.