Temporal Language Modelling of Biomedical Text: A Novel Text Mining Approach for Biomarkers Prediction

TECHNISCHE UNIVERSITÄT DRESDEN

biotec Biotechnology Center TU Dresden

Orestis Gkorgkas, George Tsatsaronis, Iraklis Varlamis, Kjetil Nørvåg

HUA-DIT
Harokopio
University of
Athens

NTNU
Norwegian University of
Science and Technology

Abstract

The increasing volume of biomedical literature. e.g., PubMed indexed articles constitutes a huge data source for applying text mining and predicting trends and new biomedical terminology. In this work we explore, for the first time to the best of our knowledge, the application of temporal language models in the PubMed indexed literature, in order to identify trends in terms. The suggested methodology comprises three steps: (i) training of temporal language models using a parametric window of time, (ii) application of the generated temporal language models to unseen scientific literature in order to mine the properties of the underlying terms and classify the literature in time windows, and, (iii) extraction of new biomedical terms that are suggested by the temporal models as terms following an ascending trend, and which may play a very important role in the future, e.g., biomarkers.

Background and Motivation

Many MESH ontology terms appear in very few MEDLINE documents and therefore they have an important distinguishing role in the document collection. We can exploit this feature and retrieve the most important terms that characterize a period. We can therefore use these terms to estimate the date a document was written based on its content.

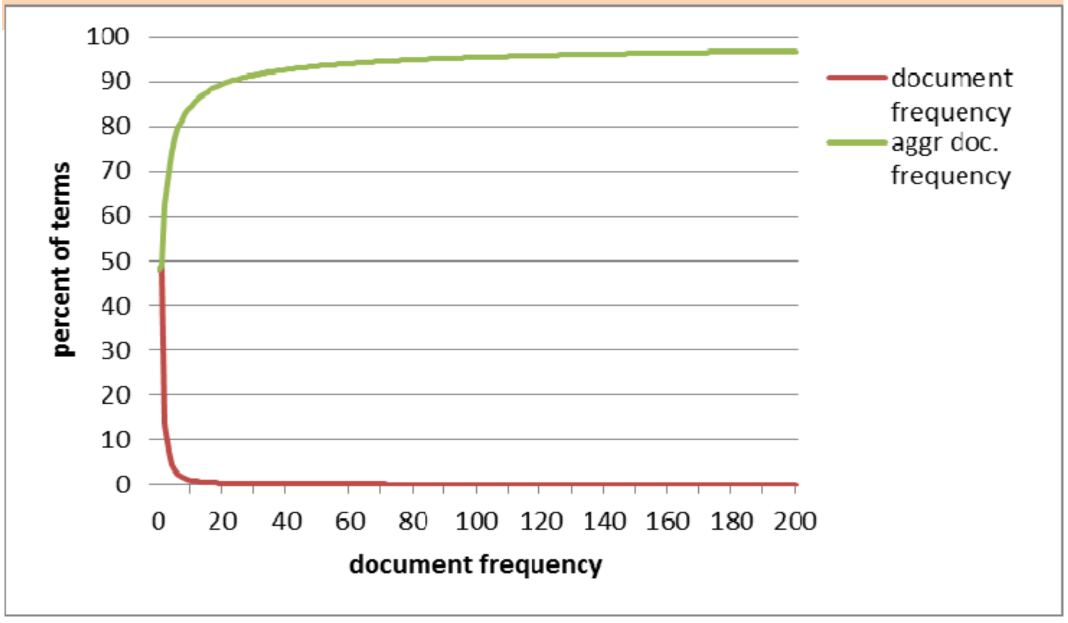

Figure 1: Distribution and aggregate distribution of the document frequencies of all terms in MESH document collection. Almost 90% of the terms appear in at most 20 documents in a period of 42 years (1970-2011).

Methodology

In this early stage we follow an adopted version of the methodology described in [1]. Our goal is to verify if we can estimate accurately enough the period a medical paper was published. In order to achieve that we build a positive and a negative temporal model for each time period.

The positive temporal model of a period is a term-vector which consists of all the terms that appear during that period. However we do not use terms that appear only once in the whole document collection since they do not provide any useful information. In that way to reduce the number of different terms being processed about 50%. The negative temporal model consists of all or some of the periods that precede the current period.

Each document is tested against the positive and the negative model of a period and we for each model a score is produced. If the positive score is larger than the negative one, it is assumed that the document belong to the period. The score is calculated using Formula 1. Formula 2 describes the temporal entropy of a term as that is defined in [1].

$$Score_{te}(d_i, p_j) = \sum_{w \in d_i} TE(w) \times P(w|d_i) \times \log \frac{P(w|p_j)}{P(w|C)} \qquad (1)$$

$$TE(w_i) = 1 + \frac{1}{\log N_P} \sum_{p \in \mathbf{P}} P(p|w_i) \times \log P(p|w_i) \qquad (2)$$

Experimental Results

During the experiments we focused mainly on two parameters. The size of the negative model (look-back depth) and the length of the periods (time period length). The length of the periods is actually the accuracy window within which we are trying to estimate the date of a document. As shown in the graphs of Figure 2, the length of the model does not play a significant role in estimating the date of the document. We can therefore use small negative models, consisting of few past periods.

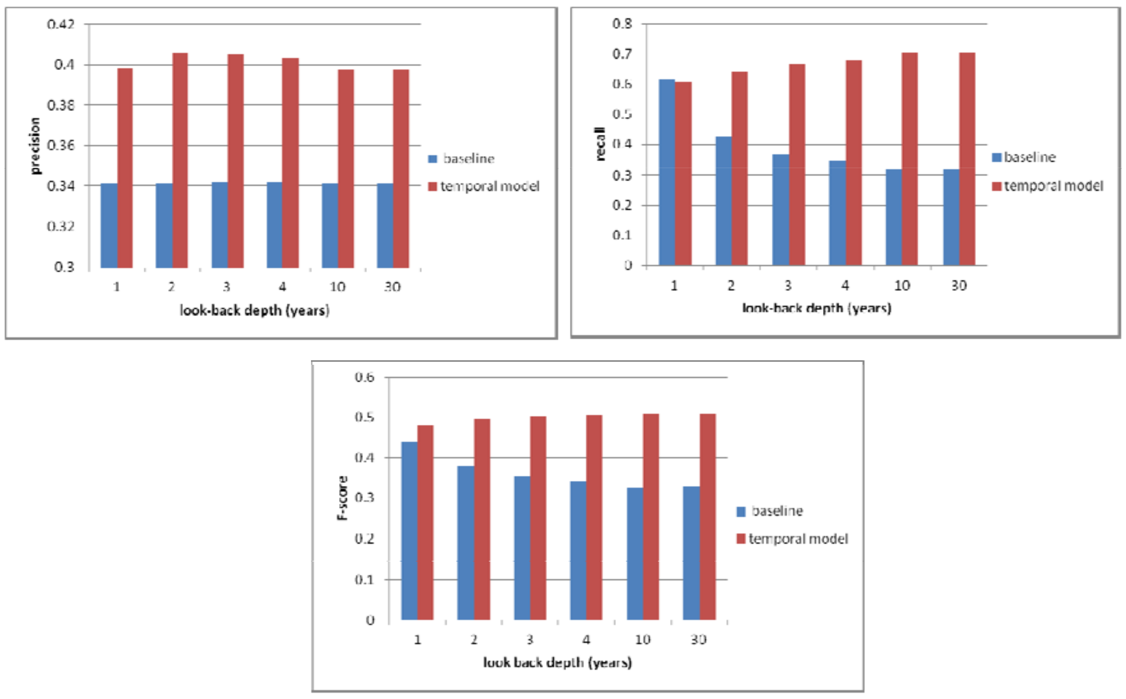

Figure 2: Precision, Recall and F-score values for different sizes of the training set (look-back depth).

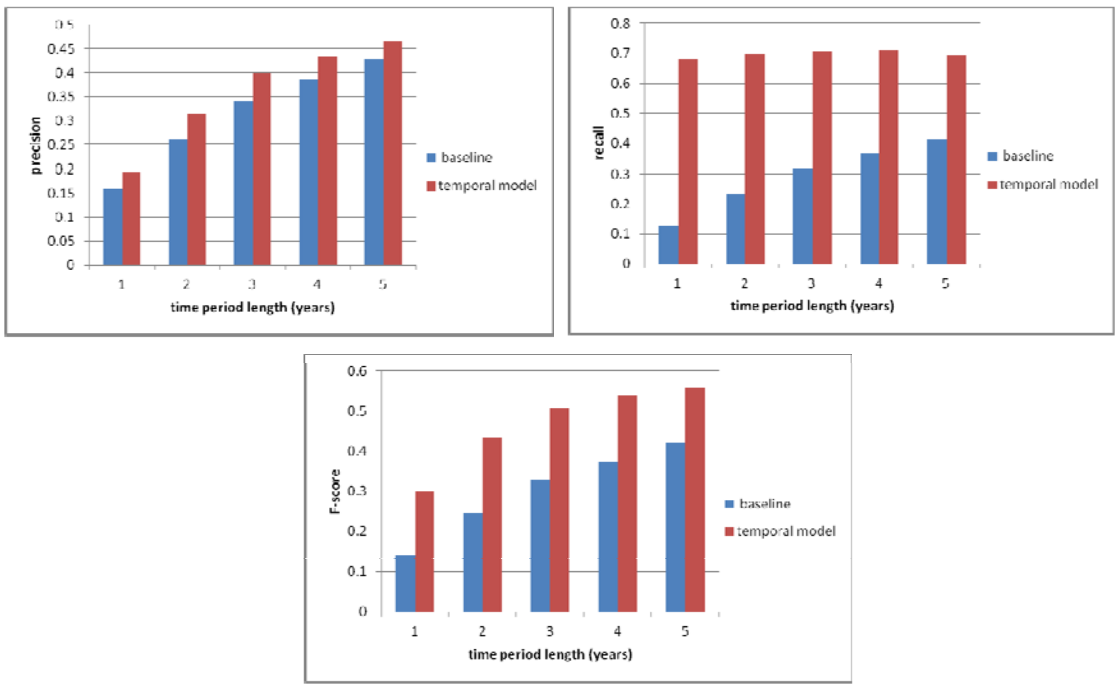

Figure 3: Precision, Recall and F-score values for different lengths of the examining dating period (time period length).

Figure 3 shows that the time period length affects heavily the precision. Estimating the dating of documents at the level of a single year is a very difficult task with poor precision. However, considering larger time frames, e.g., a time-span of five years, allows the model to predict with higher precision, and, thus, F-Score. Recall values remain unaffected and high, independently of the time period length.

Conclusions and Future Work

Initial experimental evaluation for steps (i) and (ii), considering all of the biomedical literature indexed by PubMed since 1970, shows that the suggested temporal language models can train efficiently when time windows of a three-year time span are used, and ten-fold cross validation analysis shows that we can accurately predict the time window of a scientific paper with an F-Measure of almost 77%, and by considering only title, abstract, and MeSH headings. The next steps will be to extract features that can accurately predict the new biomedical terms, given the features and predictions of the models in (i) and (ii).

References

[1] Nattiya Kanhabua, Kjetil Nørvåg: Improving Temporal Language Models for Determining Time of Non-timestamped Documents. ECDL 2008: 358-370