

The causality between cigarette and lung cancer: addressing issues with NLP

Hongcheng Jiang, Liye Zhu, Yachen Zhao, Zhaobin Zhang

1. Topic and Specialization

The topic of this proposal is exploiting the causality between cigarette and lung cancer, which has become increasingly serious recently years. More and more people get lung cancer. Based on research, lung cancer is the second most common cancer in both men and women. What is more, lung cancers can be hardly cured. Among all the causes, cigarette is the most primary reason causing lung cancer. The more cigarettes you smoke per day, the earlier you started smoking, the greater your risk of lung cancer.

Though modern medical technique has achieved significant achievement, it is still not applicable to review all the historical materials and documents for doctors to explore the hidden relationships between cigarette and lung cancer. So, we decided to collect the data of the relationships between smoking and lung cancer. Natural language processing (NLP) is adopted to process the data to investigate the latent causality. This proposal aims at exploiting the causality between smoking/cigarette and lung cancer thus providing valuable reference for current medical diagnosis and treatment.

2. Related Work

In [1], Y. H. Khang proposed that the relationship between smoking and lung cancer has to applied on multiple people. This is a key reference pointing out that discovering the hidden causality should reply on statistical techniques, not a single phenomenon. Even if smoking contributes to 80%-90% of lung cancer occurrence, the contribution cannot be applied to individuals because this is a statistical result observed in population group [2].

Graham W. [3] proposed that the lung cancer risks are not reduced by switching to filters or low-tar/low-nicotine cigarettes. The reason behind that is more complicated than expected. In patients with cancer, continued tobacco use after diagnosis is associated with poor therapeutic outcomes including increased treatment-related toxicity.

3. Dataset and Ontology Resources

PubMed API is adopted as the primary dataset in this proposal, which is free biotechnology resources. In the PubMed API, we can find some theses about our topic. Practically, we want to find the relationship between Cigarette and lung cancer. Hence, we choose the paper as our dataset.

- [1]. Young-Ho Khang, The causality between smoking and lung cancer among groups and individuals: addressing issues in tobacco litigation in South Korea, [online] doi: 10.4178/epih/e2015026, 2015 May 31.
- [2]. Preparatory pleadings for KT&G, Phillip Morris International Korea, British American Tobacco Korea BAT in the compensation claim 2014GaHap525054 (Korean).
- [3]. Graham W. Warren, Tobacco and Lung Cancer: Risks, Trends, and Outcomes in Patients with Cancer, [online] 10.1200/EdBook_AM.2013.33.359, 2013 ASCO Annual Meeting.