Automatic Knowledge Extraction for News Articles

Mengtian Jin[1], Yilong Li[2], and Linying Yang[1]

[1]Institute for Computational and Mathematical Engineering, Stanford University
[2]Department of Computer Science, Stanford University
{mtjin, yilong, lilyy}@stanford.edu

I. Motivation

A structured knowledge database of news articles is very important in a variety of areas like news agencies, media websites and some knowledge-based decision making systems. It can be used to recommend customized news articles for users, analyze development of a business by synthesizing similar articles. In this project, we aim at building an interface that will use NLP to automatically relate articles and knowledge by topic. We aim to have a system that integrate news into a predefined knowledge base, based on a database of tagged news articles, and to identify key entities, topics and even high level relationships between articles automatically is important.

II. Tasks

Before integrating news articles into a predefined knowledge base, there are some basic tasks for the news dataset, and these are what we will do in this project:

A. General Topic Recognition

We should be able to get the topic information from the text itself, which is usually the category of the news should belong to. There can be multiple tags for a single article, for example, the article *US judge orders temporary pause on deportations of reunited migrant families* can be related to tags like "Immigration" "Law & Justice" and "The Facts".

B. Named Entity Extraction and Linking

For news categorization we also need to **extract the key named entities** in the article and divide them into different categories like *Places*, *Regions*, *Persons* and *Groups*. There are usually a lot of named entities included in this article but what we need is only the key entities for article tagging. Also for one word or phrase, it can refer to different named entities based on its contexts. So it will be also important to **link the named entity** we extracted to the semantic unit itself.

III. Methods

A. Named Entity Extraction

Our group will first experiment on entity extraction. Given the fact that NER field has been thriving for more than twenty-five years. Although most of the work has concentrated on limited domains and textual features, good techniques are available in terms of news articles.

Handcraft rules and machine learning algorithms can classify techniques of entity extraction. Typical rule-based systems like SystemT [1] and GATE [2] make use of rules throughout the entire flow. Due to many special cases, handcraft rules are not sufficient and time-consuming. Supervised learning approaches such as HMM, Decision Trees, SVM and ME are current dominant techniques. However, the crucial dependence on the availability of large, representative and high-quality labeled training datasets makes building large-scale NER systems hard. Thus, researches on applications of semi-supervised learning and unsupervised learning without the prerequisite of an annotated corpus are more popular these days and we will try different techniques in this part.

B. Named Entity Linking

Another important experiment will be done is entity linking. Words can be ambiguous. A same word can refer to different things. Therefore we want to link the word to the entities that it is truly referring to in our knowledge base. Entity linking problem can be formulated as a ranking problem. If we vectorize the descriptions in a document to a feature vector x, and let y be a target entity that is in set Y, a set of all possible entities of x, we can then create a scoring function, $\hat{y} = f(y, x)$ [3]. SVM will be an applied to train the scoring function by minimizing a margin-based ranking loss.

C. Topic Recognition

For this project, the topic recognition process can be regarded as a multi-label classification, where multiple labels may be assigned to a single instance. In the dataset labels can be correlated and data may be unbalanced [4]. To solve the topic recognition problem, we can either transform the problem to single-label problems, or adapt existing single-label learning methods to multi-label cases. Per [5], there are several state-of-the-art multi-label classification algorithms, including transformation methods like chained classifiers [6], or adaptation methods like Rank-SVM [7], decision-tree based ML-DT [8] and modified k-nearest neighbor algorithm ML-KNN [9]. We can use some of these methods as a baseline to start with and then improve our topic recognition metrics.

IV. Database

For the news, we will use the database provided by The File LLC, which is a large CSV file containing the title, contents, and human-labelled information like topics, regions, events, persons, groups, places and themes for all news articles collected from the news agency during a long time period. And

we are going to use the **DBpedia** [10] which contains a large number of entries, each of which has a paragraph of abstract, descriptions and links to related DBpedia types and entries, as our knowledge base for integration.

V. Intended Experiments

For all the problems, we can divide our news data into three datasets: **training set** for training the model, **validation set** for tuning parameters and details, and **testing set** for final evaluation. Since we divide the project into different relatively individual parts, we can do different experiments and use different metrics to evaluate these methods.

A. Topic Recognition

For multi-label classification problem, we have a lot of metrics which can be divided into two basic types [5] : **example-based metrics**, focusing on tagging accuracy of each article, and **label-based metrics**, focusing on the label-wise statistics for each tag.

For example-based metrics, we can use metrics like example-based accuracy, precision, recall and F-value, as well as hamming loss and ranking based metrics to evaluate our algorithms. For label-based metrics, we usually calculate the true/false positive/negative values to evaluate the binary classification metrics for each class and then do some averaging over classes.

B. Named Entity Extraction and Linking

According to [11], when evaluating the named entity extraction systems we usually use the precision, recall and F measures as the metrics to evaluate the named-entity systems. We can give more weights to precision if we need to remove all unneeded entities or give more weights to get more entities collected.

References

[1] L. Chiticariu, R. Krishnamurthy, Y. Li, S. Raghavan, F. R. Reiss, and S. Vaithyanathan, "Systemt: an algebraic approach to declarative information extraction," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 128–137.

[2] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva, "Getting more out of biomedical documents with gate's full lifecycle open source text analytics," *PLoS computational biology*, vol. 9, no. 2, p. e1002854, 2013.

[3] J. Eisenstein, "Natural language processing," 2018, draft, https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf.

[4] E. Gibaja and S. Ventura, "A tutorial on multilabel learning," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, p. 52, 2015.

[5] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[6] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, p. 333, Jun 2011. [Online]. Available: https://doi.org/10.1007/s10994-011-5256-5

[7] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Advances in neural information processing systems*, 2002, pp. 681–687.

[8] A. Clare and R. D. King, "Knowledge discovery in multi-label phenotype data," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2001, pp. 42–53.

[9] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[10] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, ser. ISWC'07/ASWC'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 722–735. [Online]. Available: http://dl.acm.org/citation.cfm?id=1785162.1785216

[11] M. Marrero, S. Snchez-cuadrado, J. M. Lara, and G. Andreadakis, "Evaluation of named entity extraction systems," in *Advances in Computational Linguistics, Research in Computing Science*, 2009, pp. 41–47.