Automatic creation of metadata/markup by use of natural language processing of full text articles

First name Last name
student number
email

March 17, 2016

overview

We're producing a program which automatically generate metadata such as authors' name, date of publishing, name of articles... and more importantly auto-abstract for full text articles We are interested in features for either more convenient use of the program or improving precise data generation There are 5 related problems.

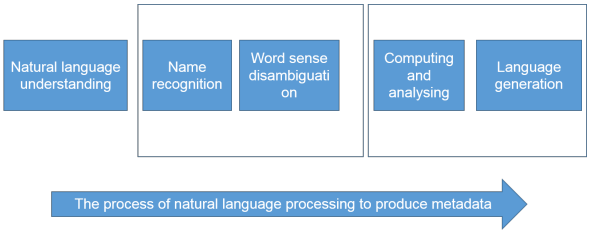

Figure 1. Overview of metadata creation process.

Problem 1

When users search with a sentence, how do the program understand the certain input of text?

Method 1

Building a natural language understanding (NLU) system.

Solution 1

Use a set of possible yes-no questions that can be applied to data items, then follow a rule for selecting the best question at any node on the basis of training data, which has a method for pruning trees to prevent over-training.

Problem 2

When users search Turkey, the results could be a country or an animal. Sometimes, the results are totally unrelated.

Method 2

categorize the results based on different subjects or genres

Solution 2

It's significantly crucial for search engine to understand what users want by name recognition in natural language processing. Digital libraries and web resources have limited metadata, augmenting them with meaningful, stable and desired categories. Information can enable better overviews and support user exploration.

Problem 3

Word sense disambiguation

Method 3

Word sense disambiguation is an important step in natural language processing. This is the step where words with different meaning will be listed in different category (Abualhaija and Zimmermann, 2016). WSD has been done with three main approaches: supervised disambiguation (Abualhaija and Zimmermann, 2016), semi-supervised approach (Ben Aouicha et al., 2016), and more recently unsupervised approach (Yoon et al., 2006). Research for unsupervised approach has been developed quickly and application of this approach has been found in WSD for not-so-popular language such as Korean.

Solution 3

The project team has decided to pursuit the unsupervised approach. Implementation will be made in term of symnonym grouping (Navigli, 2009) and context clustering (Wang et al., 2009).

Problem 4

The readers do not know what the connected words meaning

Method 4

Compute and Analyze in natural language processing

Solution 4

I suggest we should use the Python language to complete this task. It can easily conpute and analyze the words in the articles. Phthon can compute and analyze by separating the connected words. The readers can know what do the words mean when they are reading the articles .

Problem 5

Natural language processing help us extract the important information from the full text article. How could we make it more efficiently and precisely?

Method 5

Query reduction to single sub-query

Solution 5

The performance of the machine is better in the short query rather than long query. Thus, it is an important issue to reduce the query to many sub-query. The first is extracting the single sub-query by the existing features. Then, We combine these features to the reduction's technique. We could find that it is more efficient than just analyze the original query.

Reference

[1].Jin 2008, Effectiveness Web Search Results for Genre and Sentiment Classification
[2].Bill 2006, Categorizing Web Search Results into Meaningful and Stable Categories Using Fast-Feature Techniques
[3].Weiscbedel 2006 White Paper on Natural Language Processing
[4].Collins 2011 Natural Language Processing Machine Learning Research
[5].Manish Gupta 2015, Information Retrieval with Verbose Queries, Foundations and Trends in Information Retrieval
[6].Julia Hirschberg 2015, Advances in natural language processing, Science
[7].Shapiro1982, A knowledge engineering approach to natural language understanding
[8].Kuhn1995, The Application of Semantic Classification Trees to Natural Language Understanding
[9].Abualhaija, S.and Zimmermann, K.-H. 2016. D-Bees: A novel method inspired by bee colony optimization for solving word sense disambiguation. Swarm and Evolutionary Computation, 27, 188-195.
[10]. Ben Aouicha, M., Hadj Taieb, M. A.and Ezzeddine, M. 2016. Derivation of "is a" taxonomy from Wikipedia Category Graph. Engineering Applications of Artificial Intelligence, 50, 265-286.
[11]. Navigli, R. 2009. Word sense disambiguation: A survey. ACM Computing Surveys (CSUR), 41, 10.
[12]. Wang, H., Missura, O., Gärtner, T.and Wrobel, S. 2009. Context-based clustering of image search results. In: KI 2009: Advances in Artificial Intelligence. Springer.

[13]. Yoon, Y., Seon, C.-N., Lee, S.and Seo, J. 2006. Unsupervised word sense disambiguation for Korean through the acyclic weighted digraph using corpus and dictionary. Information Processing and Management, 42, 710-722.